Authors

Ana Carolina Amado Britto
Ana Luiza Marques
Caroline Alves Martins Pires Corrêa
Cristiano de Souza Therrien
Fabro Boaz Steibel
Felipe Cittolin Abal
Giovanni Simão Triginelli
Grace Li
Guilherme Ramos de Morais
Gulinigaer Wumiti
Gustavo Salema Marques
Gustavo Tadeu Zaniboni
Heitor Pedrosa Haddad
Helena Penna Marchi
Jingzhou Sun
Jiyu Zhang
Juan José Martínez Mercadal
Laura Nahabetián Brunet
Leonardo Andrade Macedo
Lucas Carvalho Corrêa
Lucas Martins de Freitas Junior
Luiz Felipe de Freitas Cordeiro
Maksim Ledenev
Marina Moretzsohn Chust Trajano
Matheus Antes Schwede
Natan Jamil Angelo
Nicolo Zingales
Richardson Xavier Brant
Rodrigo da Silva Ferreira
Rodrigo Soares Peva
Teresa Scassa
Thiago Guimarães de Barros Cobra
Yufei Ma

Leonardo **PARENTONI**
Coordinator

Ana Luiza **Marques**
Leonardo Andrade **Macedo**
Luiz Felipe de Freitas **Cordeiro**
Marina Moretzsohn Chust **Trajano**
Richardson Xavier **Brant**
Editorial Support Team



LAW, TECHNOLOGY AND INNOVATION

vol. 8: AI Regulation Landscape

DTiBR

Artificial intelligence is reshaping the foundations of contemporary legal systems, challenging established regulatory paradigms and demanding new approaches to governance, accountability, and the protection of fundamental rights. As technological development accelerates across jurisdictions, the central question is no longer merely whether artificial intelligence should be regulated, but how legal institutions can develop coherent, effective, and adaptable frameworks capable of responding to its evolving complexities.

Law, Technology and Innovation – Volume VIII: AI Regulation Landscape brings together twenty-four chapters by thirty-three contributors, offering a comprehensive and internationally oriented examination of the emerging legal and institutional architecture of artificial intelligence regulation.

Organized into six thematic parts, the volume explores the conceptual and technical foundations of artificial intelligence, international governance initiatives, comparative regulatory developments across the Americas, Europe, Asia, and the Pacific, and the intricate relationship between algorithmic technologies, fundamental rights, and systemic risks. Particular attention is devoted to regulatory experimentation, institutional oversight, digital sovereignty, constitutional guarantees, intellectual property, judicial applications, and electoral integrity.

Moving beyond purely descriptive accounts of legislation, the contributions critically examine the assumptions, limitations, and practical consequences of competing regulatory models, revealing both the possibilities and the tensions inherent in governing technologies that transcend traditional territorial, institutional, and disciplinary boundaries.

Combining theoretical sophistication with comparative analysis and attention to contemporary regulatory challenges, this collection offers scholars, legal practitioners, policymakers, and researchers a valuable framework for understanding the present landscape of artificial intelligence regulation and the transformations that will shape its future.

DTiBR

creative commons
BY SA

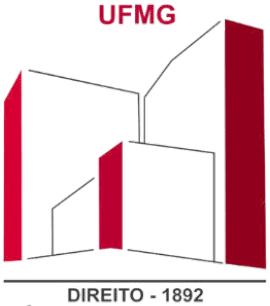
9 786599 837043

LAW, TECHNOLOGY AND INNOVATION

vol. 8: AI Regulation Landscape

DTiBR

FACULDADE DE DIREITO
UFMG



DIREITO - 1892
CIÊNCIAS DO ESTADO - 2008

DTiBR

creative commons

A presente obra foi publicada sob os direitos da Creative Commons 4.0 BY-SA. Mais informações em:
https://creativecommons.org/licenses/by-sa/4.0/

Authors

Ana Carolina Amado Britto
Ana Luiza Marques
Caroline Alves Martins Pires Corrêa
Cristiano de Souza Therrien
Fabro Boaz Steibel
Felipe Cittolin Abal
Giovanni Simão Triginelli
Grace Li
Guilherme Ramos de Morais
Gulinigaer Wumiti
Gustavo Salema Marques
Gustavo Tadeu Zaniboni
Heitor Pedrosa Haddad
Helena Penna Marchi
Jingzhou Sun
Jiyu Zhang
Juan José Martínez Mercadal
Laura Nahabetián Brunet
Leonardo Andrade Macedo
Lucas Carvalho Corrêa
Lucas Martins de Freitas Junior
Luiz Felipe de Freitas Cordeiro
Maksim Ledenev
Marina Moretzsohn Chust Trajano
Matheus Antes Schwede
Natan Jamil Angelo
Nicolo Zingales
Richardson Xavier Brant
Rodrigo da Silva Ferreira
Rodrigo Soares Peva
Teresa Scassa
Thiago Guimarães de Barros Cobra
Yufei Ma

Leonardo **PARENTONI**
Coordinator

Ana Luiza **Marques**
Leonardo Andrade **Macedo**
Luiz Felipe de Freitas **Cordeiro**
Marina Moretzsohn Chust **Trajano**
Richardson Xavier **Brant**
Editorial Support Team

LAW, TECHNOLOGY AND INNOVATION

vol. 8: AI Regulation Landscape

DTiBR

Law, Technology and Innovation – v. 8:
AI Regulation Landscape

Coordination Leonardo Parentoni

Editorial Counsel Brazilian Journal of Law, Technology and Innovation

Editorial Support Team Ana Luiza Marques
Leonardo Andrade Macedo
Luiz Felipe de Freitas Cordeiro
Marina Moretzsohn Chust Trajano
Richardson Xavier Brant

Cover Equipe Editorial

Typesetting Equipe Editorial

Centro de Pesquisa em Direito, Tecnologia e Inovação – Centro DTIBR
CNPJ/MF nº 32.727.924/0001-80
Rua dos Timbiras, 1925, Suite 903, Lourdes, Belo Horizonte, Minas Gerais, Brazil · Postal code 30140-069
www.dtibr.com

All rights reserved.

Edition closed on Oct. 2026.

Cataloging-in-Publication Data (CIP)

L415
2026

Law, Technology and Innovation – vol. VIII: AI Regulation Landscape / Leonardo Parentoni [coordinator]; Ana Luiza Marques, Leonardo Andrade Macedo, Luiz Felipe de Freitas Cordeiro, Marina Moretzsohn Chust Trajano, Richardson Xavier Brant [editorial support team]. Belo Horizonte: Centro DTIBR, 2026.
900 p.

Includes bibliographical references.
Collected volume. Multiple contributors.
ISBN: 978-65-998370-4-3
DOI: https://doi.org/10.59224/dti8

1. Law. 2. Artificial intelligence – Law and legislation. 3. Artificial intelligence – Regulation and governance. 4. Law, technology and innovation. 5. Comparative law. I. Parentoni, Leonardo. II. Marques, Ana Luiza. III. Macedo, Leonardo Andrade. IV. Cordeiro, Luiz Felipe de Freitas. V. Trajano, Marina Moretzsohn Chust. VI. Brant, Richardson Xavier.

CDU: 340/CDD: 342.2

Catalogação na fonte

Preface

I am pleased to introduce this new volume in the series "Law, Technology and Innovation", which is the first Brazilian collection to consistently evaluate a new technology every year from both a computational and legal perspective. If the reader is interested in previous volumes, the collection is available for download on the website of the Research Center for Law, Technology and Innovation in Brazil (Centro DTIBR - www.dtibr.com), under the "books and papers" section. You can also find it on Professor Perantoni's ResearchGate profile (https://www.researchgate.net/profile/Leonardo-Parentoni).

This 8th volume presents an overview of the worldwide AI Regulation Landscape, not just mentioning the core regulations and their main provisions – something that any GenAI tool can easily accomplish nowadays – but instead developing a critical analysis of the merits and shortcomings of each regulation, both computationally and legally, as well as a comparative assessment. It is an important and timely issue. AI is a pervasive technology reshaping almost every aspect of human life, from our work to our studies to how we date and watch the news. Furthermore, it impacts the economy and the way the State is governed, including sensitive areas such as warfare. In one phrase: AI may shift power.

Despite the commonality of some problems and goals in AI throughout the world, the strategies for achieving those goals may differ significantly across jurisdictions. As an example, the US heavily supports market-driven approaches, since the current global leaders are American companies: OpenAI and Anthropic. There are also other tech giants investing in AI: Nvidia, Microsoft, Alphabet (Google), Amazon, and Meta. Therefore, the American strategy relies on this unique market structure. This was especially true under the administration of Donald Trump.

There are some major AI companies in the EU, such as Mistral and Hugging Face, but they do not rival their American counterparts. Hence, the Union created a uni-

fied regulatory framework (the AI Act) to harmonize fundamental values protection and innovation, in order to establish itself as a global leader in AI. That option came with costs. In fact, the EU Parliament noted only a few years after the AI Act's enactment that the Union faces the risk of falling behind other jurisdictions because of an excess of bureaucracy. So much so that a hot topic nowadays is how to ease the burden on local developers, perhaps through amendments to the AI Act and the General Data Protection Regulation (GDPR).

Other jurisdictions adopt different strategies. China, for instance, rejected the EU's model for a unified framework. The country is also less market-oriented than the US. In fact, China developed a network of multiple parallel approaches, comprising ethical norms, the adjudication of regular civil and administrative laws alongside new legislation focused on specific issues, such as the deep synthesis regulation. It is all supplemented by self-regulation and mandatory technical standards. It appears this complex arrangement is working, since China is challenging the US's protagonism in AI by taking the lead in patent filings and volume of scientific publications.

While drawing their national AI strategies, Australia, Brazil, Canada, Russia, Uruguay, and other countries openly considered these and other sources. However, the results were different in each country. The perspectives and solutions to similar problems may also differ across international frameworks from institutions such as OECD, UNESCO, and more recently WAICO.

Summing up: there is no one-size-fits-all solution to AI regulation. It is natural that different jurisdictions will experiment with different techniques which are best suited to their particular cultural, economic, and political traditions, while also keeping an eye on successful initiatives from abroad.

Hence, this book can be useful for a wide audience. Students taking their first steps in the field may be able to unravel the big picture of AI regulation, while senior professionals may gain insight into the topic in greater depth. Additionally, the book was carefully written to assist not only legal professionals, such as practicing lawyers or judges, but also scholars in economics, computer science, social and health sciences, whenever they are interested in better understanding AI regulation.

33 coauthors, representing 7 nationalities, contributed to the following pages. With their diverse backgrounds in computer science, history, information systems,

law, mathematics, political communication, and public administration, they provide a multidisciplinary view that goes beyond strict legal analysis.

The book is structured in 06 Parts. Part I discusses the conceptual, technical, and theoretical foundations of AI and their relevance to legal regulation. Starting with Gustavo Zaniboni's provocative piece, which questions the "AI system that does not exist". Afterwards, Lucas Martins presents a critical analysis of the EU AI Act governance structure, emphasizing a software-agnostic approach. Then Rodrigo Ferreira discusses AI regulation trends, uncertainty, and the value of layered models.

Part II examines international and transnational regulatory initiatives, soft law instruments, and global governance mechanisms. A discussion of digital sovereignty and technology dependency in some countries, such as Brazil, takes place between Cristiano Therrien and Felipe Abal. Next, Fabro Steibel, Natan Angelo, and Ana Carolina Britto examine the second wave of AI governance institutions. While Thiago Cobra compares AI regulatory sandboxes worldwide.

Part III of the book covers comparative studies of AI regulation in the Americas and Europe. Teresa Scassa explains Canada's strategy; A broad range of Latin American countries is covered by Rodrigo Peva (Chile, Colombia, Costa Rica, Ecuador, Mexico, Panama, Peru, and Uruguay); Using the Brazilian system as a comparison point, Heitor Haddad analyzes the new legislative framework of the EU; In their article, Caroline Corrêa and Lucas Corrêa provide a critique and comparison of ISO/IEC 42001, the Brazilian draft law on artificial intelligence, and the EU AI Act.

Part IV details Asia and the Pacific. Grace Li opens the topic by exploring the Australian role in Asia–Pacific AI governance. Following that, Leonardo Macedo highlights the soft law foundation, safeguards, and guardrails of Australia's artificial intelligence regulation. The complex Chinese AI regulation was thoroughly examined by Jingzhou Sun, Jiyu Zhang, and Yufei Ma. Maksim Ledenev concludes the topic by reviewing the Russian legislation and drawing critical comparisons with national and international approaches.

In Part V, the authors examine how artificial intelligence impacts fundamental rights, democratic institutions, and societal values. It begins with Gulinigaer Wumiti's study on the history of US national security and the investment screening mechanism. After that, Laura Brunet examines digital constitutionalism, algorithms,

and the reconfiguration of fundamental rights. Helena Marchi then turns to the study of deepfakes, while Guilherme Morais addresses automated case law.

Part VI looks at the regulation of AI and its use in specific fields. Nicolo Zingales and Gustavo Marques deal with the thorny issue of copyrights for content generated by artificial intelligence. Juan José Mercadal examines the equally controversial issue of civil liability for damages caused by AI conversational agents to children and adolescents. Next, Richardson Brant surveys the main uses of AI in the Brazilian judiciary. Luiz Felipe Cordeiro and Marina Trajano then address the "human reviewer fallacy" by questioning the extent, scope, and feasibility of human supervision on AI-based products. The last but not least article by Ana Luiza Marques discusses the massive use of artificial intelligence in Brazil during its elections.

Other important jurisdictions could have also been addressed. For instance, the UK and India. Some factors, however, prevented us from doing so. The conclusion that AI regulation is a rapidly changing landscape was perhaps the most decisive. Therefore, our decision to proceed with the contributions that comprise this book was based on our belief that they are diverse, comprehensive, scientifically grounded, and therefore able to provide a useful global panorama on AI regulation. After all, "sometimes greater isn't better".

As a final note, it should be stated that no part of this book was written by AI. All contributions are the result of human effort on the part of the authors, and all merits are theirs alone. In an era of AI-generated content of questionable quality, this kind of disclaimer is more important than ever.

Belo Horizonte/MG, Brazil. November 2026.

Leonardo Parentoni

PhD in Law. Associate Professor at UFMG Law School/Brazil. Global Coordinator of the BRICS Law Schools Consortium Working Group on AI and Law. Former member of the original composition of the National Council for Personal Data Protection and Privacy (CNPD), appointed by the President of the Republic. Founder and Scientific Advisor to the Research Center on Law, Technology and Innovation, DTIBR (www.dtibr.com).

Institutional Support

Centro de Direito, Tecnologia e Inovação - Centro DTIBR

Centro DTIBR is a nonprofit, multidisciplinary research association that connects academia and the private sector through professional training, knowledge dissemination, and the development of innovative products and services in law and technology. It is committed to rigorous scholarship and high ethical standards.

More information: https://www.dtibr.com/;

Facebook: https://www.facebook.com/CentroDTIBR/;

Instagram: https://www.instagram.com/centrodtibr/;

LinkedIn: https://pt.linkedin.com/company/centrodtibr.

Biographies

Coordinator

Leonardo Parentoni

With more than 20 years of experience in the public and private sectors, he holds a PhD in Law from the University of São Paulo (USP), an LLM in Business Law from the Federal University of Minas Gerais (UFMG), a postgraduate specialization in Civil Procedure from the University of Brasília (UnB), and an undergraduate degree in Law from UFMG. He is a Senior Federal Attorney at the Office of the Attorney General of Brazil (AGU), an Associate Professor at UFMG Law School, and a Full Professor at IBMEC/MG. He serves as global coordinator of the BRICS Law Schools Consortium Working Group on AI and Law. He was a member of the inaugural composition of the National Council for Personal Data Protection and Privacy (CNPD), appointed by the President of Brazil. He founded and coordinates the Law, Technology and Innovation research area within UFMG Law School's Graduate Program and is founder and Scientific Council member of Centro DTIBR (www.dtibr.com). He has served on committees of Brazil's National Council of Justice, Federal Justice Council, Federal Attorney General's Office, and Brazilian Bar Association (OAB/MG). He has been a visiting researcher at the University of Texas at Austin (USA) and Uruguay's data protection authority, a strategic technology partner at the University of Technology Sydney, and a team mentor in Law Without Walls (LWOW/USA). His scholarship has been cited in decisions of Brazil's higher courts, including STF RE 1.010.606 and STJ REsp 1.925.959, and published in eight countries. His research covers (1) Law, Technology and Innovation, (2) Corporate Law, and (3) Empirical Legal Studies.

ResearcherID: N-5627-2015.

ORCID: 0000-0002-3593-2831.

Publications: https://www.researchgate.net/profile/Leonardo-Parentoni

Editorial Support Team

Ana Luiza Marques

Ph.D. Candidate and Master of Laws, Federal University of Minas Gerais (UFMG), Brazil. Visiting Researcher, Dublin City University (DCU), Ireland. Lawyer.

Leonardo Andrade Macedo

PhD candidate in Law at Universidade Federal de Minas Gerais (UFMG). Master of Laws (LL.M.) and LL.B. (UFMG). Federal Prosecutor in Brazil and Data Protection Officer of the Federal Public Prosecutor's Office (MPF). Certified Information Privacy Manager (CIPM/IAPP). Member of the Special Unit for Personal Data Protection of the National Council of the Public Prosecutor's Office (CNMP). Lecturer at Escola Superior do Ministério Público da União (ESMPU).

Luiz Felipe de Freitas Cordeiro

Master of Laws (LL.M.) from Faculdade Milton Campos

Marina Moretzsohn Chust Trajano

Master's candidate in Law, Technology and Innovation at the Federal University of Minas Gerais (UFMG)

Richardson Xavier Brant

Appellate Judge at Tribunal de Justiça de Minas Gerais, Brazil. Professor of Law and Social Development at Universidade Estadual de Montes Claros (UNIMONTES). Executive Director of the Desembargadora Jane Silva School of the Judiciary, and editor of Revista Amagis Jurídica.

Contributors

Ana Carolina Amado Britto

LL.B. (USP). Junior researcher in Law and Technology at ITS Rio. Scheduled to begin graduate studies in Commercial Law at USP in 2026.

Ana Luiza Marques

Ph.D. Candidate and Master of Laws, Federal University of Minas Gerais (UFMG), Brazil. Visiting Researcher, Dublin City University (DCU), Ireland. Law-

yer.

Caroline Alves Martins Pires Corrêa

Master of Laws, Universidade Federal de Minas Gerais. Postgraduate degrees in Procedural Law and Digital Law and Data Protection. Corporate lawyer and Information Security Analyst at NetLex.

Cristiano de Souza Therrien

Lecturer at Université de Montréal Faculty of Law and Toronto Metropolitan University. Associated researcher at University of Ottawa. LL.D., Université de Montréal.

Fabro Boaz Steibel

PhD in Political Communication (University of Leeds). Executive Director of ITS Rio. Researcher affiliated with Berkman Klein Center, Harvard University.

Felipe Cittolin Abal

Professor at Universidade de Passo Fundo. PhD in History, Universidade de Passo Fundo.

Giovanni Simão Triginelli

Master of Laws from Universidade Federal de Minas Gerais; postgraduate specialization in Health Law.

Grace Li

Professor of Law, Faculty of Law, University of Technology Sydney, Australia.

Guilherme Ramos de Morais

PhD candidate at Universidade Federal de Minas Gerais (UFMG). Master of Laws (UnB). Postgraduate Specialisation in Civil Procedure (UERJ). Civil lawyer.

Gulinigaer Wumiti

East China University of Political Science and Law.

Gustavo Salema Marques

PhD candidate at Universidade Federal do Rio de Janeiro.

Gustavo Tadeu Zaniboni

Mathematician (Unesp). Master of Science in Computer Science (UFSCar).

Founder and CAIO of Ananque Tecnologia Ltda, São José dos Campos, Brazil.

Heitor Pedrosa Haddad

Master's student in Law at Universidade Federal de Minas Gerais.

Helena Penna Marchi

Master of Laws, Universidade Federal de Minas Gerais (UFMG). Lawyer.

Jingzhou Sun

Professor of Digital Law, Renmin University of China.

Jiyu Zhang

Professor of Digital Law and Executive Dean of the Future Rule of Law Institute, Renmin University of China.

Juan José Martínez Mercadal

Attorney and specialist in Tort Law. LL.M. from Universidad de la República (Uruguay); postgraduate lecturer in Civil Law and Tort Law; PhD student at Universidad de Buenos Aires (Argentina). Managing Editor of Revista Crítica de Derecho Privado and Board Member of AIDA Uruguay.

Laura Nahabetián Brunet

Doctor of Law and Social Sciences, graduate of the Faculty of Law of the University of the Republic - Uruguay. Doctor of Legal Sciences (PhD) from the Catholic University of Argentina. Master of Science in Legislation and Political Science from the University of Pisa. Tenured Associate Professor (Grade IV) of Computer Law and Secretary of the Institute of Computer Law at the Faculty of Law of the University of the Republic. Author of more than a hundred articles in her areas of expertise. MP Advisor to the National Parliament of the Republic since 2005, currently serving with the President of the Representatives' House.

Leonardo Andrade Macedo

PhD candidate in Law at Universidade Federal de Minas Gerais (UFMG). Master of Laws (LL.M.) and LL.B. (UFMG). Federal Prosecutor in Brazil and Data Protection Officer of the Federal Public Prosecutor's Office (MPF). Certified Information Privacy Manager (CIPM/IAPP). Member of the Special Unit for Personal Data Protection of the National Council of the Public Prosecutor's Office (CNMP). Lecturer

at Escola Superior do Ministério Público da União (ESMPU).

Lucas Carvalho Corrêa

Graduated in Information Systems, Pontifícia Universidade Católica de Minas Gerais. Senior Software Engineer and Architect at LWSA Group.

Lucas Martins de Freitas Junior

Librarian, lawyer, and researcher; Master's student in Information Science at UFMG.

Luiz Felipe de Freitas Cordeiro

Master of Laws (LL.M.) from Faculdade Milton Campos.

Maksim Ledenev

Law student, National Research University Higher School of Economics (HSE); ambassador for BGP Litigation Law Firm at HSE.

Marina Moretzsohn Chust Trajano

Master's candidate in Law, Technology and Innovation at Universidade Federal de Minas Gerais.

Matheus Antes Schwede

PhD candidate in Law at Universidade Federal de Minas Gerais; Master of Laws in Human Rights from UNIJUÍ.

Natan Jamil Angelo

LL.B. (PUC-SP). Master's candidate in Philosophy and General Theory of Law, Universidade de São Paulo. Researcher at ITS Rio and research assistant at LSE.

Nicolo Zingales

Professor at FGV Direito Rio.

Richardson Xavier Brant

Appellate Judge at Tribunal de Justiça de Minas Gerais, Brazil. Professor of Law and Social Development at Universidade Estadual de Montes Claros (UNIMONTES). Executive Director of the Desembargadora Jane Silva School of the Judiciary, and editor of Revista Amagis Jurídica.

Rodrigo da Silva Ferreira

Master in Public Administration and Ph.D. Candidate in Administration at FGV/EBAPE. Attorney and Data Protection Officer at Casa da Moeda do Brasil. Lecturer and consultant on emerging technology regulation at DecisionLab.

Rodrigo Soares Peva

Master of Laws, Universidade do Estado do Rio de Janeiro (UERJ). Federal public servant at the Federal Regional Court of the Second Region (TRF2).

Teresa Scassa

Canada Research Chair in Information Law and Policy, University of Ottawa.

Thiago Guimarães de Barros Cobra

LL.M. in Administrative Law, Universidade de São Paulo. Attorney in regulatory law at HGC Advogados. Postgraduate training in civil procedure, administrative law and public management.

Yufei Ma

Postdoctoral Fellow, Renmin University of China.

Contents

Preface V
Institutional Support IX
Biographies XI
Contents XVII

I
FOUNDATIONAL CONCEPTS AND TECHNICAL GROUNDS FOR ARTIFICIAL INTELLIGENCE

CHAPTER **1** | THE AI SYSTEM THAT DOES NOT EXIST: TECHNICAL LAYERS, EFFECTIVE CONTROL, AND THE ALLOCATION OF REGULATORY DUTIES
Gustavo Tadeu Zaniboni
1. Introduction 35
2. The Anatomy of an AI-Enabled System 39
2.1 Method: NINE STAGES, THREE SYSTEMS 39
2.2 Stage One: The Corpus 40
2.3 Stage Two: Architecture and Learning Algorithm 41
2.4 Stage Three: The Training Run 43
2.5 Stage Four: Post-Training and Adaptation 44
2.6 Stage Five: The Model Artifact, or What Is Actually Delivered 46
2.7 Stage Six: Inference in Three Regimes 48
2.8 Stage Seven: The Runtime, and Why Implementation is Not Neutral 51
2.9 Stage Eight: Orchestration, Application, and Interface 52
2.10 The Agent Harness, and The Platform That Builds It 53
2.11 Stage Nine: Deployment Topology, and Why The Interface Dominates 55
2.12 The Cost Gradient and Its Structural Consequence 56
3. Control, Composition, and Where Risk Becomes Knowable 57
3.1 Three Capacities, and Why They Come Apart 57
3.2 The Determination Gradient: Capability Upward, Purpose Downward 59
3.3 Reversibility as An Allocative Criterion 60

3.4 Orthogonality: AI-Intensity Is Not Risk-Intensity 60
3.5 Composition, and Why Component-Wise Assessment Fails 62
3.6 The Limit of The Thesis: Capability-Intrinsic Risk 64
3.7 The Allocative Criterion, and What Follows From It 65
4. Two Illustrations 66
4.1 Facial Recognition: The Axis of Purpose 66
4.1.1 What a Facial Recognition System Actually Is 66
4.1.2 Deployments A and B: The Benign Cases 68
4.1.3 Deployment C: Watchlist Search In Public Space 69
4.1.4 What Varied, and What Did Not 71
4.1.5 What the AI Act Gets Right, and Where It Stops 72
4.2 One Model, Three Topologies: The Axis of Control 73
4.2.1 Three Configurations 73
4.2.2 What Varies 73
4.2.3 What The Law Sees 75
4.2.4 Why The Least Auditable Configuration Prevails 77
5. The Duty Of Specification and The Duty of Selection 77
5.1 An Existing Solution To The Same Problem 77
5.2 Why The Speed Rating Works 78
5.3 The Content of a Model Specification 79
5.4 Why Current Model Cards Are Not Specifications 81
5.5 Trade Secrecy, Enforcement, and Why Disclosure Must Be Mandatory 81
5.6 The Correlative Duty of Selection 83
5.7 The Limits of The Analogy 84
5.8 The Framework Applied 85
5.9 The AI Act As Enacted 87
Conclusion 89
References 92

CHAPTER **2** | ARTIFICIAL INTELLIGENCE GOVERNANCE ARCHITECTURES: A CRITICAL ANALYSIS OF THE EUROPEAN AI ACT, THE SOFTWARE-AGNOSTIC APPROACH, AND A HYBRID REGULATORY MODEL BASED ON RISKS AND EVIDENCE

Lucas Martins de Freitas Junior

1 Introduction 96
2 The Crisis In Defining Artificial Intelligence: Between Technology, Law, and Regulation 98

3 Artificial Intelligence as a Distinct Technology or Simply Another Form of Software ?..101
4 The European Union's Horizontal Paradigm: The AI Act And Risk-Based Governance 105
5 The Effectiveness Problem: Predictive AI, Technological Hype, and The AI Snake Oil Phenomenon 109
6 A Proposal For Hybrid Regulatory Governance of Artificial Intelligence 113
6.1 First Layer: General Principles Applicable To All Digital Systems 114
6.2 Second Layer: Regulation Based on The Context of Use 115
6.3 Third Layer: Specific Requirements For Artificial Intelligence Systems 115
6.4 Fourth Layer: Evidence-Based Governance 116
6.5 A Regulatory Architecture Guided By Proportionality 116
7 The Construction of Artificial Intelligence Governance In Brazil: Between International Influence and The Specific Features of The National Legal System 117
7.1 The Federal Constitution as the Foundation of Artificial Intelligence Governance 118
7.2 The General Data Protection Law and the Consolidation of Digital Governance 119
7.3 The Brazilian Artificial Intelligence Strategy and The Construction of a National Regulatory Framework 120
7.4 The Role of Public Institutions In Constructing Artificial Intelligence Governance ... 121
7.5 Challenges to Consolidating Artificial Intelligence Governance In Brazil 122
7.6 Brazilian Artificial Intelligence Governance In Comparative Perspective 123
8 Conclusions 124
References 125

CHAPTER **3** | FROM AI SYSTEMS TO AUTOMATED ACTIVITIES: FUTURE TRENDS IN ARTIFICIAL INTELLIGENCE REGULATION
Rodrigo da Silva Ferreira
1. Introduction 127
2. The Limits of AI-Centered Regulation 132
2.1. Technological Risk and the Regulatory Problem of Uncertainty 132
2.2. AI as a Problematic Regulatory Object 135
3. Toward Activity-Based Governance 138
3.1. From AI Systems to Sociotechnical Activities 138
3.2. A Layered Model for Future Regulation 141
3.3. Building Institutional Capacity 145
4. Conclusion 148
References 150

II
INTERNATIONAL GOVERNANCE AND REGULATION OF ARTIFICIAL INTELLIGENCE

CHAPTER 4 | BEYOND THE CLOUD ACT: DIGITAL SOVEREIGNTY, TECHNOLOGICAL HEGEMONY AND INSTITUTIONAL PATHWAYS FOR REGULATION OF ARTIFICIAL INTELLIGENCE IN BRAZIL
Cristiano de Souza Therrien; Felipe Cittolin Abal
Introduction 158
1 Digital Sovereignty and Technological Hegemony: Two Concepts In Tension 161
2 The Cloud Act, Fisa Section 702 and The Legal Architecture of Extraterritorial Access 170
3 Brazilian Digital Dependence: Manifestations, Risks and Institutional Responses 177
3.1 The São Paulo Court of Justice and Microsoft 177
3.2 DATASUS and the health data of the Brazilian population 180
3.3 From Latent Risk To Open Confrontation: The 2025–2026 Sanctions and Trade Offensive 182
3.4 Institutional Responses: The Sovereign Cloud, The AI Bill and The Data Centre Policy 185
3.5 Why Compute, Not Just Data 189
Conclusion 191
References 196

CHAPTER 5 | FROM AISI TO AISO: BRAZIL ON THE SECOND WAVE OF AI GOVERNANCE INSTITUTIONS
Fabro Boaz Steibel; Natan Jamil Angelo; Ana Carolina Amado Britto
1. Introduction 206
2. AISIs and The First Wave 208
2.1 What are AISIs 208
2.2 The AISIs' Origin 210
2.2.1 UK AI Safety Summit 211
2.2.2 AI Seoul Summit 213
2.2.3 AI Action Summit (Paris) 215
2.2.4 India AI Impact Summit 216
2.3 AISIs - First wave 217
2.3.1 United Kingdom 217
2.3.2 United States of America 218
2.3.3 Singapore 220

2.3.4 Canada 221
2.3.5 Japan 222
2.3.6 France 223
2.3.7 European Union 225
2.3.8 South Korea 226
3. From AISI to AISO: The Brazilian Proposal 227
4. An AISO in Brazil: Characteristics and Functions in Practice 228
4.1 Characteristics 233
4.2 Functions 236
5. Conclusion 239
Bibliography 240

CHAPTER **6** | AI REGULATORY SANDBOXES BEYOND THE HYPE: COMPARATIVE DESIGN CHOICES, GOVERNANCE, AND METRICS FOR ACCOUNTABILITY
Thiago Guimarães de Barros Cobra
1. Introduction 246
2. Theoretical Framework: Regulation, Risk, and Institutional Learning 249
2.1. Regulating Emerging Technologies and The Limits of The Prescriptive Model 249
2.2. Risk-Based Regulation: Promise, Blind Spots, and Systemic Risk 251
2.3. Responsive and Experimental Regulation: Institutional Learning as a Regulatory Strategy 253
3. International Normative Backdrop: The Foundations of "Trustworthy AI" 255
3.1. European Union: from "Trustworthy AI" To Normative Densification 257
3.2. United Kingdom: Sectoral Coordination, and Adaptability Instruments 259
3.3. United States: Institutional Fragmentation, and Political Volatility 260
3.4. Asian Experiences (China, South Korea, and India): State-Centered Models 262
3.5. Digital Sovereignty: Data, Infrastructure, and Governance Tensions 265
4. Institutional Design of AI Sandboxes: Governance, Metrics, and Limits 267
4.1. Purposes of AI Sandboxes: Regulatory Learning and Supervised Compliance 269
4.2. Eligibility and Selection: Scope Boundaries and Risk-Proportionate Design 270
4.3. Safeguards and Risk Management Throughout The Experimental Cycle 271
4.4. Governance, Continuous Oversight, and Incident Documentation 273
4.5. Exit Rules and Institutionalizing Learning: From Testing to Decision 274
5. Metrics, Evidence, and The Evaluation of Experimentalism In AI 276
5.1. Performance, Robustness, and Operational Security: Minimum Evidence For a Reliable AI Test 277

5.2. Socioeconomic Impacts and Distributive Effects: Benefits, Costs, and Externalities of The Experiment 278
5.3. Evidence Governance and Cycle Integrity: Reporting, Auditability, and Comparability of Results 279
6. When The Sandbox Becomes a Dangerous "Shortcut": Risks and Conditions of Legitimacy 281
6.1. Exceptionalism Becomes the Rule: Cycle Perpetuation and Erosion of Exit Rules 282
6.2. Competitive Asymmetry and "Sandbox Shopping": Regulatory Arbitrage and Market Distortions 283
6.3. Opacity and Capture: Informational Asymmetry and The Weakening of Public Learning 284
Conclusion 285
References 286

III
COMPARATIVE AND GLOBAL PANORAMA: AMERICAS AND EUROPE

CHAPTER 7 | CANADA'S MIDDLE POWER AI REGULATION
Teresa Scassa
1. Context 293
2. A Federal State With Obligations to Indigenous Communities 297
2.1 Federalism 298
2.2 Significant Indigenous Communities 302
3. Sovereignty and AI 305
3.1 Infrastructure 306
3.2 Data 309
3.3 Policy 313
4. AI Regulation in Canada 315
4.1 Federal law/policy 318
4.1.1 Federal Public Sector 318
4.1.2 Private Sector 321
4.2 Provincial law/policy 324
5. Conclusion 328
References 329

CHAPTER **8** | AI REGULATION IN LATIN AMERICA
Rodrigo Soares Peva
1. Introduction 345
2. Overview of AI Regulation 347
3 Criteria For Categorizing Latin American Regulatory Instruments 352
4 Regulatory Mapping in Latin America 353
4.1 Argentina 353
4.2 Brazil 356
4.2.1. Brazil's National AI Plan – PBIA 356
4.2.2. Brazil's multilevel AI governance 358
4.2.3. Subnational Lawmaking 359
4.2.4. Sectoral Governance: Judiciary, Elections, and Criminal Law 361
4.2.5 A Broadened Conception of AI Sovereignty 362
4.3 Chile 363
4.4 Colombia 364
4.5 Costa Rica 365
4.6 Ecuador 367
4.7 Mexico 368
4.8 Panama 371
4.9 Peru 373
4.10 Uruguay 376
5. 2023 Santiago Declaration 378
6. What Can Be Learned From The Latin American Regulatory Trend ? 379
Conclusion 380
References 383

CHAPTER **9** | THE NEW LEGISLATIVE FRAMEWORK (NLF) AND THE DANGER OF DECONTEXTUALIZED IMPORTATION OF THE ARTIFICIAL INTELLIGENCE ACT TO BRAZIL
Heitor Pedrosa Haddad
1. Introduction 392
2. Methodology 393
3. Old Approach x New Approach 393
4. New Legislative Framework (NLF) 397
4.1. Regulation 765/2008, Regulation 2019/1020 and Decision No. 768/2008/EC 399
4.2. Evaluation of the New Legislative Framework 405

5. Artificial Intelligence Act X General Data Protection Regulation 407
5.1. Artificial Intelligence Act 408
5.2. General Data Protection Regulation 411
6. Brazilian Bill 2,338/23 415
7. The Product Safety System in Brazil 418
8. Conclusion 424
References 425

CHAPTER **10** | ISO/IEC 42001, DRAFT BILL NO. 2.338/2023 AND EUROPEAN AI ACT: A CRITICAL-COMPARATIVE ANALYSIS OF GOVERNANCE AND REGULATION OF ARTIFICIAL INTELLIGENCE IN BRAZIL
Caroline Alves Martins Pires Corrêa; Lucas Carvalho Corrêa
1. Introduction 430
2. Technical Standard - Artificial Intelligence ISO/IEC 42001 (Soft Law) 431
3. Europe - AI Act (Hard Law) 433
4. Draft Bill No. 2.338/2023 (Hard Law) 436
5. Technical Comparison: ISO/IEC 42001 vs PL 2.338/2023 and EU AI Act 437
5.1 Risk Management 438
5.2 Organizational Governance 439
5.3. Ethical Principles 441
5.4. Other Obligations and Duties 442
5.5. Prohibited Systems 445
5.6. Sanctions 447
5.7. Public Sector 448
5.8. Table of Contents 449
Conclusion 454
References 455

IV
COMPARATIVE AND GLOBAL PANORAMA:
ASIA AND PACIFIC

CHAPTER **11** | BETWEEN FRAGMENTATION AND CONVERGENCE: AUSTRALIA'S ROLE IN ASIA–PACIFIC AI GOVERNANCE
Grace Li
1. Introduction: Australia and The Asia–Pacific AI Governance Moment 460

2. The Asia–Pacific AI Governance Landscape: A Spectrum of Approaches 463
2.1 Principles-Based and Guidance-Oriented Models 464
2.2 Hybrid Innovation–Regulation Frameworks 466
2.3 Centralised and Prescriptive Regulatory Approaches 467
2.4 Regional Challenges and Fragmentation 468
3. Australia's AI Governance Architecture 469
3.1 Sectoral Regulation as De Facto AI Law 473
3.2 Strengths and Limitations of Australia's Approach 475
4. Comparative Perspectives: Australia and Key Asia–Pacific Jurisdictions 476
4.1 Australia and Japan: Normative Alignment, Institutional Divergence 477
4.2 Australia and Singapore: Pragmatism and Operationalisation 478
4.3 Australia and South Korea: Regulatory Readiness and State Capacity 479
4.4 Australia and New Zealand: Shared Values, Different Regulatory Emphases 480
4.5 Australia and India: Developmental Priorities and Regulatory Sequencing 480
4.6 Australia and China: Contrasting Governance Logics 481
4.7 Comparative Insights 482
5. Regional Cooperation, Soft Convergence, and Australia as a Bridging Jurisdiction 483
5.1 Soft Convergence through Multilateral Dialogue 483
5.2 Standards and Functional Interoperability 485
5.3 Australia's Bridging Role in Regional AI Governance 485
5.4 Strategic Implications 486
6. Conclusion 487
References 489

CHAPTER **12** | DIVERGENT PATHS IN AI REGULATION: AUSTRALIA AND BRAZIL AT A MOMENT OF GLOBAL RECALIBRATION

Leonardo Andrade Macedo

1. Introduction 495
2. Regulating AI: Rationales and Strategies 498
3. Australia's AI Regulatory Trajectory 502
3.1 Ethical Principles as Soft-Law Foundations 503
3.2 Consultation, Risk, and The Case For Mandatory Safeguards 507
3.3 Voluntary Guardrails: From VAISS to AI6 510
3.4 Mandatory Guardrails and The Technology-Neutral Pivot 512
4. Australia and Brazil In Comparative Perspective: Divergent Paths, Mutual Lessons 519
5. Concluding Remarks 524

References 525

CHAPTER 13 | REGULATING ARTIFICIAL INTELLIGENCE IN CHINA: A TIERED, AGILE AND EMPOWERING GOVERNANCE MODEL

Jingzhou Sun; Jiyu Zhang; Yufei Ma

1. Introduction: Beyond The Quest For a Unified Artificial Intelligence Act 532
2. Normative Foundations of China's Artificial Intelligence Governance 535
(I) Human-Centeredness and Beneficial AI 535
(II) Balancing Development and Security 538
(III) Inclusive and Prudent Regulation Coupled with Agile Governance 539
(IV) Capacity Building and Sustainable Development 542
3. The Tiered Architecture of China's Artificial Intelligence Regulation 544
(I) Ethical Norms and Science-and-Technology Ethics Review 544
(II) General-Purpose Laws as the Fundamental Protective Layer 546
(III) Targeted Special-Purpose Regulation for Specific Technologies and Risks 548
(IV) Sector-Specific Regulation and Scenario-Based Governance 550
(V) Local-Level Experiments, Industrial Promotion and International Governance 552
4. Technical Standards: The Bridge Linking Legal Principles and Compliance Practice 554
(I) Dual Governance Demands for Artificial Intelligence and Functional Complementarity between Law and Standards 554
(II) Manifestations and Underlying Causes of Insufficient Coordination 556
(III) Soft-Law Attributes of Technical Standards and the Transformation Path of "Legal Principles-Technical Standards-Compliance Tools" 558
(IV) Dynamic Coordination Mechanisms: Regular Communication, Sandbox Supervision and International Cooperation 560
5. Pluralistic, Agile and Empowerment-Oriented Governance Mechanisms 562
(I) Government Supervision: From Command-and-Control to Tiered-Classification and Continuous Interaction 562
(II) Corporate Self-Governance and Industry Collaboration 563
(III) Social Supervision, Individual Rights and Remedies 565
(IV) Empowering Innovation and Compliance-Building Capacity 568
6. Existing Challenges and The Global Significance of The Chinese Model 569
(I) Norm Fragmentation and Hierarchical Coordination 569
(II) Further Improvement Needed for Public Participation and Right Remedies 570
(III) Legitimacy of Technical Standards and Risks of Regulatory Capture 571
(IV) Insights from China's Experience for the Global South 571

Conclusion 572
References 573

CHAPTER **14** | ARTIFICIAL INTELLIGENCE IN CHINA: THE REGULATORY MOVEMENT AND ITS RESPECTIVE STAGES OF DEVELOPMENT
Matheus Antes Schwede; Giovanni Simão Triginelli
1 Introduction 576
2 Methodology 577
3 The First Stage of Artificial Intelligence Regulation in China 579
4 The Second Stage of Artificial Intelligence Regulation in China 588
5 Considerations on The Chinese Regulatory Model 594
References 602

CHAPTER **15** | PROTECTION OF RIGHTS OF ARTIFICIAL INTELLIGENCE: COMPARATIVE PERSPECTIVES ON ASSIGNING LEGAL RESPONSIBILITY IN RUSSIAN AND GLOBAL PRACTICE
Maksim Ledenev
Introduction 608
Review of Sources 610
2. Analysis of Regulatory and Other Official Acts of The Russian Federation 612
2.1. Review of Russian Legislation 613
2.2. A Review of Official Texts on AI Development in The Russian Federation That Are Not of a Regulatory Nature 615
3. Analysis of International and Foreign Regulatory and Legal Acts on AI 616
3.1. The Approaches Proposed in European Union 616
3.2. Approaches Proposed in The USA 619
3.3. Approaches Proposed by UNESCO 621
3.4. Approaches Proposed in Other Countries 622
4. Review of Judicial Practice 624
4.1. Analysis of international practice in resolving AI disputes 624
4.2. Analysis of trends in AI dispute resolution 625
5. Possible Approaches to Protecting AI Rights 627
5.1. Foundations for The Development of a Special Model For Regulating The Protectability of AI in The Russian Federation 627
Conclusion 630
References 631

V
ARTIFICIAL INTELLIGENCE: BETWEEN FUNDAMENTAL RIGHTS AND SYSTEMIC RISKS

CHAPTER **16** | THE SECURITIZATION OF ARTIFICIAL INTELLIGENCE: NATIONAL SECURITY EXPANSION AND ITS INSTITUTIONAL IMPACT ON INVESTMENT SCREENING
Gulinigaer Wumiti
Introduction 640
1. The Five Transformations of The U.S. National Security Concept 641
1.1 The First Transformation: The Cold War and the Focus on Ideological Threats 641
1.2 The Second Transformation: The Nixon Era and the Balance of Power 644
1.3 The Third Transformation: The Clinton Era and the Rise of Economic Security 645
1.4 The Fourth Transformation: 9/11 and the Shift to Preemption 647
1.5 The Fifth Transformation: Dual-Use Tech and the Weaponization of National Security 649
2. Current Landscape of U.S. Investment Screening Mechanism 650
2.1 Inbound Investment Screening: The CFIUS Framework 650
2.2 Outbound Investment Screening: The OISP Mechanism 656
3. Analysis of The Drivers Behind The Tightening of AI Investment Scrutiny 660
4. Assessment of The Tightened U.S. Investment Review System 663
Conclusion 665
References 666

CHAPTER **17** | ALGORITHMIC CONSTITUTIONALISM: FROM THE GUARANTEE TO THE RECONFIGURATION OF FUNDAMENTAL RIGHTS
Laura Nahabetián Brunet
1. Classical Constitutionalism and Its Historical Evolution 670
2. Algorithmic Power as a Constitutional Challenge 671
3. Algorithms as Instruments of Power 672
4. Digital Governance and Algorithmic Constitutionalism 673
5. Algorithmic Governance 677
6. Ethical Principles for a Technological World 679
7. From The Guarantee to The Reconfiguration of Fundamental Rights 680
7.1. The Transition from The Individual to The Collective 682
7.2. Fundamental Rights in The Age of Algorithms 683
7.2.1. Privacy and Data Protection: From Limit to Design Principle 684

7.2.2. Equality and Non-Discrimination: Towards Algorithmic Equity 685
7.2.3. Freedom of Expression and Media Diversity in Digital Environments 685
7.2.4. Human Dignity and Autonomy in The Algorithmic Age..... 686
8. Constitutional Culture in The Digital Age 686
8.1. The Hybrid Nature of The Contemporary World 687
8.2. Modernising the Constitution..... 689
8.3. Algorithmic Constitutionalism and Democratic Legitimacy 690
8.4. Algorithmic Power and Its Implications for Democracy..... 691
8.5. Algorithmic Constitutionalism as a Paradigm of Legitimacy 691
9. Conclusions..... 692
References..... 693

CHAPTER **18** | REGULATING DEEPFAKES IN BRAZIL: INSIGHTS FROM INTERNATIONAL EXPERIENCE
Helena Penna Marchi
1. Introduction 695
2. Deepfakes: Definition and Characteristics 698
3. Potential Harms Caused by Deepfakes and Grounds For Regulation 703
4. Deepfakes Regulation 707
4.1. Deepfakes Regulation in China 707
4.2. Deepfakes Regulation in The European Union 709
4.3. Deepfakes Regulation in The United States 713
4.4. Regulatory Models in Summary..... 715
5. Paths for Regulation of The Matter in Brazil..... 716
6. Conclusion 719
References..... 720

CHAPTER **19** | CODEX: PROSPECTS AND CHALLENGES IN THE AUTOMATED ANALYSIS OF CASE LAW
Guilherme Ramos de Morais
1. Introduction 726
2. Planning, Development, and Implementation of Codex 727
3. Technical Features of Codex..... 730
4. Discussion on The Potential Impacts of Codex on Improving Information Retrieval ... 733
4.1. Unified National Indexing..... 734
4.2. Nationwide Semantic Retrieval 740

5. Regulatory Challenges Related to The Implementation of Codex and Its Integration with AI Models 742
5.1. Data Access and APIs 742
5.2. Information Security 744
5.3. Personal Data Protection 747
6. Final Remarks 748
References 750

VI
ARTIFICIAL INTELLIGENCE IN REGULATED SECTORS AND THE DIGITAL ECONOMY

CHAPTER **20** | COPYRIGHT FOR THE INPUTS OF ARTIFICIAL INTELLIGENCE: BALANCING FREE ENTERPRISE WITH THE RIGHTS OF AUTHORS
Nicolo Zingales; Gustavo Salema Marques
1. Introduction 758
2. Generative AI Training and The Copyright Problem 762
2.1 Expressive Works and the Generative AI Supply Chain 764
3. Regulatory Responses to Text and Data Mining 769
3.1. Global Regulatory Overview 769
3.2. U.S. and South Korea's Fair Use and its application in Brazil 778
3.2.1. What is fair use and how does it works 778
3.2.2. Comparative Overview of U.S. and South Korean Fair Use Provisions 780
3.2.3. Comparative Overview of U.S. and South Korean Fair Use Guidelines 782
4. Fair Use as a Regulatory Strategy for AI Training in Brazil 792
5. Conclusion 795
References 796

CHAPTER **21** | CONVERSATIONAL AI, CHILDREN, AND ADOLESCENTS: NAVIGATING THE FRONTIERS OF CIVIL AND PRODUCT LIABILITY
Juan José Martínez Mercadal
1. Introduction: Dystopia or Reality? 802
2. The Zoon Politikon in Crisis: The Rise of Conversational Agents 804
3. The Displacement of Human Trust Toward AI: Implications for The Emotional Development of Children and Adolescents 807
4. Once More on Vulnerability in AI Times: Is It An Inherent Condition? 809

5. Children's Rights in The Digital Environment: General Comment No. 25 as a Framework 812
6. Are Conversational Agents Defective Products? Lessons From Recent Us Common Law 814
6.1. *Megan Garcia v. Character Technologies, Inc* 814
6.2. *P.F., et al. (KGM) v. Meta Platforms, Inc.* 820
7. Final Reflections on Civil and Product Liability 822
References 828

CHAPTER **22** | REGULATING THE USE OF GENERATIVE ARTIFICIAL INTELLIGENCE IN THE JUDICIARY: NOTES ON THE ETHICAL STANDARDS OF SAFETY AND TRANSPARENCY
Richardson Xavier Brant
1. Introduction 832
2. Surveys on The Use of AI in The Judiciary: Benefits and The Partnership with The UNDP 835
3. The Safety and Transparency Standards Laid Down By The CNJ 839
4. Proper Use By Judicial Professionals: Judges and Court Staff 842
5. Risks and The Prevention of Harm 845
6. Conclusion 847
References 848

CHAPTER **23** | THE "HUMAN REVIEWER" FALLACY: AN ANALYSIS OF HUMAN SUPERVISION IN ARTIFICIAL INTELLIGENCE SYSTEMS IN THE JUDICIARY
Luiz Felipe de Freitas Cordeiro; Marina Moretzsohn Chust Trajano
1. Introduction 852
2. The Need For Normative Differentiation Among AI Applications According to Risk Profiles In The Judiciary 854
3. The Use of Large Language Models (LLMs) to Support Decisions on The Merits 857
3.1 Opacity and The "Black Box" Problem 859
3.2 Algorithmic and Cognitive Biases in Artificial Intelligence 861
3.3 Automation Bias and the Anchoring Effect 866
4. Conclusion 869
References 869

CHAPTER **24** | SECTORIAL REGULATION OF ARTIFICIAL INTELLIGENCE IN ELECTIONS: THE CASE OF BRAZIL

Ana Luiza Marques
Introduction 874
1. Artificial Intelligence Regulation: General Approaches and Sectorial Strategies 876
2. Social Media Platforms and Artificial Intelligence: Transnational Challenges to Democracy and Elections 882
3. The Brazilian Response: Resolution nº 23.732/2024 From The Superior Electoral Court 890
3.1 Contextual Background: Electoral Institutions And Disinformation Challenges 890
3.2 Resolution nº 23.732/2024: Regulatory Innovations and Enforcement Mechanisms . 894
3.3 Comparative Perspectives: International Approaches to AI and Electoral Integrity ... 897
Conclusion 902
References 904

I
FOUNDATIONAL CONCEPTS AND TECHNICAL GROUNDS FOR ARTIFICIAL INTELLIGENCE

THE AI SYSTEM THAT DOES NOT EXIST: TECHNICAL LAYERS, EFFECTIVE CONTROL, AND THE ALLOCATION OF REGULATORY DUTIES

Gustavo Tadeu Zaniboni

Mathematician (Unesp). Master of Science in Computer Science (UFSCar). Founder and CAIO of Ananque Tecnologia Ltda, São José dos Campos, Brazil.

DOI: https://doi.org/10.59224/dti8.ch1

Abstract: Global instruments converge on one regulated object, the AI system. This chapter argues that the object does not exist. Disaggregating an AI-enabled system into its stages shows that capability is fixed upstream and purpose downstream, so context-dependent risk becomes knowable only at the application layer, while knowledge and control rest with different parties; capability-intrinsic risks remain assessable upstream. It proposes a duty of specification upstream and a duty of selection downstream, and shows that this dissolves the general-purpose AI category.

Keywords: Artificial intelligence regulation; effective control; specification duty.

Table of contents: 1. Introduction. 2. The anatomy of an AI-enabled system. 3. Control, composition, and where risk becomes knowable. 4. Two illustrations. 5. The duty of specification and the duty of selection. 6. Conclusion. References.

1. INTRODUCTION

The expression *AI system* has become the fundamental unit of artificial intelligence regulation. Article 3(1) of the European AI Act defines it and erects an entire

risk apparatus upon it.[1] The OECD Recommendation supplied the formulation that most jurisdictions went on to adopt.[2] Brazil's proposed general framework follows the same architecture.[3] Technical bodies, working from different premises and toward different ends, converged on compatible vocabulary, and although terminology standards and voluntary risk frameworks do not themselves allocate duties, their alignment has helped fix the impression that a single object is out there waiting to be governed.[4] The convergence has produced real gains in regulatory interoperability, and this chapter does not dispute them.

It argues that the object does not exist as a stable engineering object capable of bearing a coherent set of duties, and that the difficulty is not merely one of technical inaccuracy. An entity that cannot be identified cannot bear a duty. Regulation built on the AI system directs obligations at a subject incapable of receiving them, with the predictable consequence that they settle on whoever finds them cheapest to document rather than on whoever could have affected the outcome.

What regulators call an *AI system* is not an engineering object at all. It is a temporary composition of independent layers: a training corpus, an architecture, a training run, post-training adaptation, a set of weights, an inference runtime, orchestration logic, retrieval components, external tools, an interface, human operators, and a hosting arrangement. Each is produced at a different moment, by a different party, under different constraints, and each can be assembled and reassembled without notice to anyone.

The instability of that composition is not incidental to the argument; it is the argument. No single layer determines the behavior of the resulting application. One model may power thousands of systems that have nothing in common, and one application may replace its underlying model, or abandon learned components

---

1 Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence, art. 3(1), 2024 O.J. (L 1689).

2 OECD, *Recommendation of the Council on Artificial Intelligence*, OECD/LEGAL/0449 (adopted 22 May 2019, amended 3 Nov. 2023).

3 *Projeto de Lei* No. 2338/2023 (Braz.).

4 ISO/IEC 22989:2022, *Information Technology — Artificial Intelligence — Artificial Intelligence Concepts and Terminology*; NAT'L INST. OF STANDARDS & TECH., *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*, NIST AI 100-1 (2023).

altogether, without any change in what it does or in how it presents itself to the people it affects. That last possibility deserves emphasis, because it is the *reductio* of the whole approach: an application can leave the scope of every instrument surveyed above while its function, its risk profile, and its effect on the people subject to it remain exactly what they were. Coverage, on this account, is something an engineer can switch off.

A concrete case sharpens the abstraction. A bank scoring consumer credit with a gradient-boosted ensemble falls squarely within the high-risk regime; its board, weighing compliance cost against a marginal loss of predictive power, replaces the ensemble with a decision tree whose variables, splits, and thresholds are set by the credit committee itself. The applicant completes the same form, receives the same refusal, and has the same recourse or lack of it. What changed is the provenance of the tree's parameters: fitted, it is machine learning and the system is regulated; authored, it is ordinary software and the system is not.[5] Nor is the authored tree necessarily the safer artifact, since a fitted model is at least disciplined by a validation set, while a tree whose thresholds were chosen in a committee room encodes that room's priors directly and answers to nothing. The boundary of the so-called AI system moves with architectural choice rather than with technological identity, which means the regulated object is not stable even within a single deployment over the course of its own life. A conformity assessment of such an object certifies a snapshot.

Two consequences follow. The first can be shown in a sentence. Consider a payroll application that uses one logistic regression to flag anomalous expense claims: it is an AI system under every definition above, and its provider is therefore exposed to whatever obligations the applicable regime attaches to that status. Consider next an industrial inspection platform in which learned components perform nearly every operation, deployed in a closed facility where a technician verifies every output

---

[5] Regulation (EU) 2024/1689, *supra* note 1, Recital 12 (distinguishing systems that infer from traditional software operating on rules laid down solely by natural persons) and Annex III (classifying creditworthiness assessment of natural persons as high-risk). The Commission's guidelines on the definition of an artificial intelligence system, C(2025) 924 final (6 Feb. 2025), elaborate the distinction and place simple deterministic tools and basic statistical estimators outside the definition, without resolving the case of a fitted model replaced by an authored one of the same functional form.

before anything happens: it carries almost no risk at all. *AI-intensity and risk-intensity are orthogonal axes.* A category that sorts along the first while purporting to govern according to the second will over-include the innocuous and under-specify the dangerous, and no refinement of the definition repairs this, because the defect lies in the choice of subject rather than in the precision with which it was drawn.

The second is that risk is not a property of any component, the model included. It is a relation between what a component can do and the circumstances in which it is made to do it. A facial recognition model trained on the faces of Nordic men has a performance envelope: it performs well on some populations and poorly on others. That is a specification, not a defect. Controlling access to a Swedish military academy, it is fit for purpose. Identifying suspects on the streets of Rio de Janeiro, it is a discrimination engine. Same weights, same measured accuracy, same artifact, incommensurable risks. The discrimination was never in the model. It came into existence when someone chose a population.

From this the chapter derives two propositions. Risk is *compositional*: it emerges from the interaction of learned and non-learned components, and the non-learned components frequently dominate, since whether a match triggers an arrest or merely surfaces a candidate for human review is decided in ordinary application code. And context-dependent risk becomes knowable only at the application layer, the layer at which purpose, affected population, decision authority, and remedy are fixed. Every layer below determines capability. Only the application determines what that capability is for.

None of this implies that model producers should be relieved of obligation. It implies that their obligation is of a different kind, and that the difference is not one of degree. A producer cannot assess the risk of applications not yet conceived, in markets unknown, affecting populations unspecified; requiring it to perform such an assessment guarantees either paralysis or theater. What it can do, and what only it can do, is state what the artifact is: on what data it was built, with what measured performance across which conditions and which populations, and where the envelope ends. A tire manufacturer does not assess whether a given car will be driven on an unrestricted autobahn. It molds a speed rating into the sidewall.[6] The vehicle

[6] U.N. Econ. Comm'n for Eur., Regulation No. 30, *Uniform Provisions Concerning the Approval of*

constructor selects a tire adequate to the vehicle and answers for that choice; Section 5.2 sets out the four properties that make such a rating work. This is not an exotic proposal, but the component parts doctrine of products liability applied to a new class of component: a duty of *specification* upstream, a duty of *selection* downstream. [7] Current instruments do not ignore the value chain, and the AI Act plainly does not. They allocate obligations by commercial role and contractual position rather than by technical layer and effective control, so the party bearing the duty is routinely the party least able to perform it.

The argument proceeds in four movements. Section 2 decomposes an AI-enabled system into its constituent stages, tracing three concrete systems through each of them, and closes on the agent harness, where the distance between legal intuition and engineering reality is widest. Section 3 asks what each stage can and cannot determine, and establishes that knowledge and control are held by different parties at different times. Section 4 tests the analysis against two cases, one isolating the axis of purpose and one the axis of control. Section 5 sets out the two duties, confronts the objections of unspecifiability and of commercial secrecy, and locates the proposal against the AI Act as enacted.

2. THE ANATOMY OF AN AI-ENABLED SYSTEM

2.1 Method: nine stages, three systems

To reject a unified object is to incur an obligation to supply a better one. This section disaggregates an AI-enabled system into the nine stages that actually compose it, establishing for each what artifact exists, who produces it, what it costs, and what it fixes irreversibly.

Two failure modes threaten any attempt to do this in a legal text: abstraction so general that no allocation of duties can follow, and detail so fine that no jurist will read it and no regulator can operationalize it. The route between them is a worked

Pneumatic Tyres for Motor Vehicles and Their Trailers.

[7] *See* RESTATEMENT (THIRD) OF TORTS: PRODUCTS LIABILITY § 5 (AM. L. INST. 1998).

example, carried through every stage, and three hypothetical systems will serve.

Cifra is a check-processing component at a retail bank, which transcribes handwritten numerals from deposit slips so that amounts can be posted automatically. *Nordica* is a facial verification model developed by a Scandinavian vendor and licensed to integrators as a component; it is not itself a product. *Atendia* is a customer-service assistant deployed by a Brazilian electricity utility, built on a publicly released large language model and connected to the utility's billing systems.

The three were chosen because they differ along every axis that matters. *Cifra* is narrow, entirely built in-house, and carries modest risk. *Nordica* is narrow but its risk varies by orders of magnitude depending on where it is installed. *Atendia* is general in capability, assembled almost entirely from components its deployer neither built nor can inspect, and intermediate in risk.

2.2 Stage one: the corpus

Everything a learned component knows about the world, it knows because the world was sampled and the sample was presented to it. The corpus therefore fixes three things at once: the population the resulting model is able to represent, the legal provenance of the material it was built from, and the privacy exposure it carries forward.

Cifra is trained on MNIST, a public dataset of seventy thousand handwritten digits at 28px×28px resolution, derived from a larger collection assembled by the United States National Institute of Standards and Technology from census employees and secondary-school students.[8] MNIST is the most benign corpus imaginable: fully public, extensively documented, benchmarked in the literature for nearly three decades. And yet it encodes a population. Its digits reflect North American handwriting conventions; the crossed seven and the serifed one common in continental Europe and in Latin America are sparsely represented. A model trained on it will read some hands better than others. This is worth stating early because it disposes of a widespread misconception. It is not that some corpora are biased and others are

[8] Yann LeCun et al., *Gradient-Based Learning Applied to Document Recognition*, 86 PROC. IEEE 2278 (1998).

clean. *Every corpus is a sample of a population*, and what is called bias is the name given to the discrepancy between that population and the one encountered in deployment. Bias is not a contaminant to be removed. It is a coordinate, and it can only be read against a destination.

Nordica's corpus is proprietary: several million facial images collected by the vendor under consent instruments of its own design, predominantly from Northern European subjects and skewed toward adult men. Its composition is simultaneously the vendor's principal commercial asset and its principal legal exposure, and it will not be published. The integrator who licenses *Nordica* therefore acquires a component whose most decision-relevant property is undisclosed.

Atendia's base model was pre-trained by its producer on a corpus reported to exceed fifteen trillion tokens drawn from publicly available sources, with composition described only in aggregate terms.[9] Neither the utility that deploys *Atendia* nor its regulator nor its customers can establish what is in that corpus. This is not a failure of diligence by the utility. It is a structural condition of the layer.

Note the asymmetry already visible at stage one. *Cifra*'s developer knows everything about its corpus and could document it exhaustively. *Atendia*'s deployer knows essentially nothing and cannot. A documentation obligation phrased identically for both is satisfiable by one and not the other.

2.3 Stage two: architecture and learning algorithm

Two distinct choices are routinely conflated under the word "algorithm," and separating them matters. The *architecture*, or model family, specifies the functional form to be fitted: what shape the mapping from inputs to outputs is permitted to take. The *learning algorithm* specifies the procedure by which the parameters of that form are adjusted. Both are expressed as source code, and neither has any content: an architecture is not *about* anything, and the same architecture serves medical diagnosis and advertising placement without modification.

---

[9] AI@META, *Llama 3 Model Card* (2024), https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md (reporting pretraining on over 15 trillion tokens drawn from publicly available sources, with composition described only in aggregate).

The available families are more varied than public discussion suggests. Linear and logistic regression fit weighted sums of named variables. Decision trees partition the input space by thresholds on individual features; random forests and gradient-boosted ensembles aggregate many such trees. Support vector machines find separating boundaries in transformed spaces. Convolutional neural networks exploit spatial locality in images. Transformers, introduced in 2017, process sequences through learned attention over positions.[10] The choice among them is an engineering decision with legal consequences that are almost never attributed to it.

Cifra illustrates this precisely. On MNIST, a logistic regression achieves roughly ninety-two percent accuracy; a random forest reaches approximately ninety-seven; a modest convolutional network exceeds ninety-nine. The bank's engineers selected the convolutional network. In doing so they purchased slightly more than two percentage points of accuracy at the cost of native explainability: each tree in a random forest splits on an identifiable pixel threshold, so its verdict decomposes into a finite set of readable rules, whereas the convolutional network's does not decompose at all. That trade, accuracy against contestability, was made by an identifiable person at an identifiable moment, it is documented in a repository, and it determines whether a customer whose deposit is misread can ever be told why. It is exactly the kind of decision regulation ought to reach, and no current instrument does.

Nordica is a convolutional embedding model trained with a margin-based metric-learning objective.[11] Its architecture entails something of first-order importance for the argument of this chapter: *the model does not decide whether two faces match*. It maps a facial image to a vector of five hundred and twelve numbers. Whether two such vectors are close enough to count as the same person requires a threshold, and no threshold exists anywhere in the model. Selecting it is the act of choosing a position on the trade-off between false acceptances and false rejections, which is to say the act of deciding whose errors matter more. That choice is made by the integrator,

---

[10] Ashish Vaswani et al., *Attention Is All You Need*, in ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS 30 (2017).

[11] *See, e.g.*, Jiankang Deng et al., *ArcFace: Additive Angular Margin Loss for Deep Face Recognition*, in PROCEEDINGS OF THE IEEE/CVF CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION 4690 (2019).

in application code, after the model is delivered. The vendor cannot make it, because the correct threshold depends on the consequence of an error, and the consequence is a property of the deployment.

Atendia's base model is a transformer decoder trained on the objective of next-token prediction. That objective, and nothing more, is what the model was built to do.

What does a failure at this stage look like? The question deserves a direct answer, because it is often assumed that architectural error is the paradigm case of algorithmic harm, and it is not. An unsuitable architecture is one that cannot represent the task: the loss fails to descend, validation accuracy stalls, and the deficiency is apparent before deployment. Architectural failures are failures of capability and ordinarily surface in testing, where they are visible to the developer and comparatively cheap to correct. This is a point with allocative significance: the stage whose failures announce themselves is the stage that needs the least regulatory attention.

2.4 Stage three: the training run

Training is the process by which an architecture, initialized with meaningless parameters, becomes a model. For neural networks the procedure is gradient-based optimization with backpropagation, and its logic can be stated compactly. A batch of examples is passed forward through the network to produce outputs. A loss function compares those outputs to the desired ones and returns a single number measuring error. Backpropagation applies the chain rule of differential calculus to compute, for every parameter in the network, the derivative of that number with respect to that parameter, that is, the direction in which each parameter should move to reduce error. An optimizer then adjusts every parameter by a small step in that direction. The cycle repeats, for *Atendia*'s base model, on the order of millions of times.

Two features of this process are indispensable to any competent legal analysis of it.

The first is that *the loss function carries the entire normative content of training*. It is the only place in the pipeline where "better" is defined, and the model becomes whatever answers the question the loss function asked. Nothing else is optimized,

and nothing not represented in the loss is protected.

The second is that *the model is a mirror of a process in the world*. Training does not learn what is correct; it learns what is regular in the data. Where the data records a social process, what the model acquires is a representation of that process, including whatever is wrong with it. A model trained on historical arrest records to predict re-arrest has faithfully modeled policing, not criminality. A model trained on a firm's past hiring decisions has modeled that firm's past preferences, not merit. The consequence is uncomfortable and must be stated without hedging: in such cases a *less accurate* model would score worse by the only metric available at this stage, and a technically impeccable training process yields a socially unacceptable artifact. This is not a defect in execution. It is a limit of the method, arising from the choice of what to model, and that choice is invisible from inside the loss function. It follows that documentation of training-stage diligence, however rigorous, cannot certify that the modeled process was a fit object of modeling, and that any regime treating such documentation as sufficient will certify unjust models as compliant.

The economics of this stage are equally consequential. *Cifra*'s convolutional network trains in a few minutes on a single graphics processor. The producer of *Atendia*'s base model reported expending on the order of 1.3 million accelerator-hours to train its eight-billion-parameter variant and approximately 6.4 million for its seventy-billion-parameter variant.[12] The gap is not a matter of degree; it is the difference between an afternoon's work and a capital project. And the training run is a single, irreversible event. Once complete, the weights are what they are. Retraining does not patch a model; it produces a different one.

2.5 Stage four: post-training and adaptation

A model emerging from pre-training predicts continuations of text. It does not answer questions, follow instructions, or decline requests. Those behaviors are installed afterwards, by a family of techniques collectively termed post-training: supervised fine-tuning on curated demonstrations; preference optimization, whether by

[12] AI@META, supra note 9 (reporting 1.3 million and 6.4 million H100-80GB accelerator-hours for the 8B and 70B variants respectively, 7.7 million in total).

reinforcement learning from human feedback or by more direct methods; distillation of a large model's behavior into a smaller one; quantization, which reduces the numerical precision of the weights to lower memory requirements; and safety conditioning, which installs refusal behaviors.

Among these, parameter-efficient adaptation deserves particular attention from lawyers, and low-rank adaptation is its dominant form.[13] Rather than updating all of a model's parameters, the technique freezes them and inserts small trainable matrices into selected layers, training a few million parameters instead of tens of billions. The resulting adapter is a file of a few dozen megabytes, and producing one takes hours on hardware available for a few hundred dollars.

The regulatory implication should be stated as plainly as possible. *The behavioral profile of a released model, including the removal of its safety conditioning, can be materially altered by one person with commodity hardware, at a cost five or six orders of magnitude below the original training run, producing an artifact that is neither traceable nor distinguishable by name from the original*. Model identity is not stable under cheap modification. The AI Act's provision on substantial modification gestures toward this problem, and administrative guidance has since supplied a threshold expressed as a proportion of the compute used to train the original model.[14] But a compute ratio is the wrong measure for an intervention whose cost is five or six orders of magnitude below the training run: the least expensive and least traceable intervention in the entire pipeline sits, by construction, far below the line.

Atendia illustrates the point concretely. The utility trains a low-rank adapter over the eight-billion-parameter base model on twelve thousand curated support transcripts, teaching it the firm's tariff vocabulary and escalation etiquette. What is now deployed is a distinct artifact with a distinct behavioral profile, which the base model's producer did not create, did not test, cannot inspect, and has no means of learning

---

[13] Edward J. Hu et al., *LoRA: Low-Rank Adaptation of Large Language Models*, in INTERNATIONAL CONFERENCE ON LEARNING REPRESENTATIONS (2022).

[14] Commission guidelines on the scope of the obligations for providers of general-purpose AI models (July 2025), treating fine-tuning as giving rise to provider obligations where the compute used for the modification exceeds roughly one third of the compute used to train the original model. A low-rank adapter falls several orders of magnitude below that line.

about. The producer's model card describes the base. Nothing describes what the utility built. If specification is to be a legal duty, as Section 5 proposes, the question of who specifies an adapted model cannot be avoided.

Note also that the pre-training producer released two models, at eight and seventy billion parameters. Same corpus, same architecture, same training pipeline, differing only in scale. But the smaller runs on a laptop while the larger requires roughly one hundred and forty gigabytes of accelerator memory at half precision, or two datacenter-class devices at minimum. In law they are the same "model," released under the same license and described by a single card. In fact they are different objects, with different capability envelopes and incompatible deployment economics.

2.6 Stage five: the model artifact, or what is actually delivered

When a model changes hands (licensed, published, downloaded, or exposed behind an interface), what is transferred is narrower than legal discourse assumes. A trained model consists of three things: a vocabulary, or tokenizer, mapping fragments of text to integer identifiers; a configuration file recording architectural and hyperparameter choices; and a set of weight matrices, which for a contemporary language model contain billions of floating-point numbers. Nothing else. There is no corpus, no index, no record set, no retrieval mechanism, and no table in which training material sits waiting to be consulted.

This bears emphasis because a persistent assumption in policy debate holds that a model contains the data on which it was trained, and that questions of erasure and data subject rights can be resolved by locating and removing the relevant entries. There are no entries. Training does not deposit texts into a store; it adjusts numerical parameters so as to reduce prediction error over a corpus, and the corpus is thereafter, in a strict engineering sense, absent from the result. What remains is a large numerical object shaped by exposure to data it does not hold. That such an object can generate fluent prose about matters it does not store, without ever operating on a whole word, is the closest thing in the field to *mathemagic*, and the term is used advisedly: the mechanism is entirely understood at the level of the arithmetic and

remains genuinely unintuitive at the level of the result.[15]

Two clarifications are required, and both matter legally.

The first is that the absence of a database does not entail perfect amnesia. Models demonstrably reproduce sequences from their training data under some conditions, and extraction of such sequences has been achieved experimentally.[16] Memorization is real. But it is not storage. It is a statistical property that emerges under identifiable conditions (data duplication, model scale, sequence rarity), and it is therefore measurable and, to a substantial degree, mitigable through deduplication and training-time intervention. The correct legal characterization is neither that a model is a copy nor that it is unrelated to its training data, but that it exhibits a *quantifiable propensity to reproduce*, which is a property that can be tested for and reported. The distinction is not academic: a duty to disclose measured memorization rates is a duty a producer can discharge, whereas a duty to purge records from a structure containing none is incoherent.

The second concerns what is actually being licensed. Weights are distributed in specific serialization formats, and conversion between them, particularly quantization to four-bit representations for consumer hardware, produces a derivative artifact that behaves differently while circulating under the same name. Licensing compounds this. The permissive-sounding licenses under which major open-weight models are released are frequently not open source licenses in the received sense: they impose acceptable-use restrictions and, in at least one prominent case, a com-

---

[15] A neologism, used here to name the effect rather than to explain it: the arithmetic inside a deep network is fully specified and reproducible, and the competence that emerges from it is not derivable from that arithmetic by inspection.

[16] Nicholas Carlini et al., *Extracting Training Data from Large Language Models*, in 30TH USENIX SECURITY SYMPOSIUM 2633 (2021) (demonstrating verbatim extraction from GPT-2, an open-weight model executed locally); Nicholas Carlini et al., *Quantifying Memorization Across Neural Language Models*, in INTERNATIONAL CONFERENCE ON LEARNING REPRESENTATIONS (2023) (measuring memorization across the GPT-Neo family and GPT-J, open-weight models trained on the publicly available Pile corpus, thereby permitting verification of extracted sequences against the actual training data). On the methodological significance of open weights, see Section 2.11 and Section 4.2.

mercial threshold above which separate permission is required.[17] The common phrase "open source model" is therefore usually a category error. These are *open-weight* artifacts under bespoke licenses. The distinction is load-bearing, because the AI Act keys certain exemptions to free and open-source licensing, and because the Open Source Initiative's own definition for artificial intelligence requires disclosures about training data that most open-weight releases do not make.[18]

Finally, and decisively for what follows: the model artifact cannot on its own do anything at all. Weights are inert. They acquire behavior only when loaded by software that decides how to feed inputs to them, how to interpret their outputs, and what to do next.

2.7 Stage six: inference in three regimes

Legal instruments treat inference as a single phenomenon. But what happens when a trained model is put to work differs so substantially across model classes that the differences are of first-order regulatory significance, above all for explainability, where obligations calibrated to one regime are unsatisfiable in another.

Classical statistical models. In a linear or logistic regression, a generalized additive model, or a decision tree, parameters carry semantics. A coefficient attaches to a named variable and expresses a magnitude of association with a direction. A split in a tree is a threshold on an identified feature. Inference is a short arithmetic procedure reproducible by hand, and the model can be printed, read, and argued about in a courtroom. Explanation here is not an approximation constructed after the fact; the model *is* its own explanation. Where a legal regime requires that a decision be justified in terms a data subject can contest, this class satisfies the requirement natively. That deserves more weight than it receives: a great many consequential automated decisions, in credit and insurance especially, are still taken by exactly these methods,

---

[17] Meta Llama 3 Community License Agreement, https://llama.meta.com/llama3/license. The license incorporates an Acceptable Use Policy by reference, prohibits use of the materials or their outputs to improve any other large language model, and imposes additional commercial terms on licensees exceeding 700 million monthly active users.

[18] OPEN SOURCE INITIATIVE, *The Open Source AI Definition — version 1.0* (2024).

and are audited less than the generative systems that attract public attention.

Neural networks. With neural networks the semantics of individual parameters dissolve. Representations are distributed: no single weight corresponds to a nameable quantity, and the features the network relies on are learned rather than specified, existing as directions in high-dimensional spaces that need not align with any human concept. Explanation becomes a research problem rather than a reading exercise, and available techniques produce approximations whose fidelity is itself contested.

One property, however, is preserved and is routinely misdescribed. A neural network is deterministic as a function: fixed weights and a fixed input define a single output, and the small residual variation observed in practice is an artifact of execution order and hardware rather than of the model. What is lost relative to the classical regime is interpretability, not determinism. The frequent conflation of the two in legal writing obscures something important, developed immediately below: where non-determinism appears in a deployed language system, it was introduced deliberately, by a software choice, at a layer above the model.

Large language models. A large language model is a neural network operated in a specialized loop, and understanding the loop is essential to allocating responsibility for anything a chat assistant appears to do.

The raw material is not the word but the *token*. Text entering the system, whether during training or at inference, is first segmented into sub-word fragments drawn from a fixed vocabulary.[19] "Regulation" may become two or three fragments; a rare proper noun may become five. Once segmented, the fragments carry no independent meaning: they are integer identifiers whose significance exists only as position in a learned vector space and in relation to surrounding tokens. Every operation the model performs occurs over these fragments. The model does not manipulate words, sentences, or propositions, and holds no representation of any of them as such.

What the model does, on each pass, is exactly one thing: given a sequence of tokens, produce a probability distribution over the next token. That is the entire

---

[19] Rico Sennrich, Barry Haddow & Alexandra Birch, *Neural Machine Translation of Rare Words with Subword Units*, in PROCEEDINGS OF THE 54TH ANNUAL MEETING OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS 1715 (2016).

function. Everything a user experiences as more than this is supplied by software surrounding the model.

Selection. The distribution must be converted into an actual token. Whether the highest-probability candidate is taken deterministically or sampled stochastically, and with what temperature and truncation, is a decoding parameter set in the inference subsystem. Variability of output is a configuration choice, not a property of the weights.

Continuation and termination. Producing more than one token requires appending the selected token to the sequence and running the model again. The loop, and the conditions under which it stops, are implemented in software.

Memory. A model has no memory of anything. In a conversational interface, the appearance of memory is produced by the application re-transmitting prior turns with each new request. If the software does not resend the first exchange, the second response cannot reflect it, not because the model forgot but because there was never anything to forget. What is marketed as "memory" is a retrieval-and-reassembly feature of the application layer, and the misnomer has consequences: retention, deletion, and disclosure obligations concerning conversational history attach to whoever built and operates that feature, not to whoever trained the weights.

Self-knowledge. Most models have no representation of their own identity. That an assistant can answer a question about what it is, who deployed it, or what it must refuse is ordinarily attributable to a system prompt, a block of text transmitted with every request, invisible to the user, stating in effect that the model is to behave as a particular named assistant under particular constraints.[20] Remove it and the behavior goes with it. The persona is an artifact of the application.

Action. A model artifact does not itself call an interface, query a database, or execute a transaction. It emits token distributions; surrounding software selects tokens, parses any structured output, and performs the action. When those tokens are formatted as a function call, it is the surrounding software that parses them and

[20] Post-training procedures can instill a degree of self-description into the weights themselves, and some producers do so deliberately; the point in the text concerns the general case and is not affected by that refinement.

performs the action, and the decision to grant that software authority to act, together with the scope of that authority, is taken at the application layer, entirely outside the model.

The cumulative effect is easily stated. *The model completes; the software decides*. Nearly every property that makes a deployed generative system consequential (that it persists information, holds a persona, acts in the world, varies in its outputs) is contributed by components the model producer neither wrote nor controls.

2.8 Stage seven: the runtime, and why implementation is not neutral

Weights are data; running them requires an implementation. This is where legal analysis most consistently loses the thread, because it assumes a single piece of software called "the AI." There is no such thing, and the implementation that trains a model is ordinarily not the implementation that serves it.

Atendia's pipeline makes the point concrete. The base model was pre-trained by its producer in a numerical computing framework, orchestrated across thousands of accelerators by a distributed training library. The utility's adapter was trained with a third library, built for parameter-efficient methods. For evaluation, the utility loaded the adapted model through a general-purpose model library. For production serving it migrated to a high-throughput inference engine designed for concurrent request batching. And the offline version distributed to field technicians' laptops runs through a desktop model runner, which wraps a compact inference library and consumes a quantized conversion of the weights.

That is at least five independent implementations of "running this model," written by five different teams, under five different licenses, with five different defect histories. And they are not behaviorally equivalent. Divergence arises from several sources at once: the *chat template*, the convention of special tokens delimiting system, user, and assistant turns, which the model was trained to expect and which the runtime must reproduce, since a runtime that applies a different template degrades behavior, often subtly rather than obviously; default sampling parameters, which differ between runners and are frequently left untouched; quantization, since a four-bit conversion is not the model that was trained; context window handling and truncation policy, which determine what the model is shown when a conversation grows long;

and ordinary numerical variation across kernels and hardware.

The conclusion is uncomfortable for any regime that attaches duties to models. *The weights under-determine the behavior*. Two deployments of nominally the same model can answer the same question differently, and neither is malfunctioning. An obligation imposed on "the model" therefore fastens onto an object that does not by itself determine the conduct being regulated.

2.9 Stage eight: orchestration, application, and interface

Only at this stage does a system acquire a purpose, an affected population, and a decision authority. Everything below determines capability; this is where capability is aimed.

For *Cifra*, the decisive artifact is a confidence threshold. The network returns a distribution over the ten digits; the application decides that where the leading candidate falls below a set probability, the item is routed to a human clerk, and otherwise posted automatically. That threshold is the entire risk control of the system, it is a handful of lines of ordinary code, and it was set by a developer with no involvement in the model's training. Raise it and the bank absorbs labor cost; lower it and customers absorb misposted deposits.

For *Nordica*, the integrator supplies the matching threshold, the composition of the reference gallery, and, most consequentially, the action that follows a match. A system that surfaces a ranked list of candidates to a trained analyst who must corroborate independently, and a system that dispatches a patrol on a single top hit, may embed identical weights. They are not the same system, and the difference was authored entirely at this stage.

For *Atendia*, this stage contributes conversation state management, retrieval over the tariff and outage databases, tool invocations against the billing system, escalation rules, and input and output filtering. It is worth counting what is learned and what is not. The adapted model is a small fraction of the deployed codebase, and a smaller fraction still of the logic that determines whether a customer's supply is scheduled for disconnection. The consequential decisions are made by conditional statements written by the utility's engineers.

The interface then determines whether human oversight is real or nominal. Whether a confidence figure is displayed to the clerk; whether the analyst sees one name or twenty ranked candidates; whether the escalation button is adjacent to the accept button. Oversight that regulation assumes and interfaces do not support is oversight that does not exist.

2.10 The agent harness, and the platform that builds it

This is not a tenth stage. It is the orchestration stage of Section 2.9 developed to its current frontier, and it warrants separate treatment for two reasons: it is where the distance between what the law imagines and what exists is greatest, and it introduces a party that no existing value chain analysis names.

The word *agent* suggests an entity that decides and acts. What exists is a loop. Its constituents are a model, unchanged and unaware that it is being operated this way; a set of tool definitions, each a description of a function the surrounding software will execute if the model emits a properly formatted request; a controller that parses the model's output, dispatches the requested call, captures the result, appends it to the context, and invokes the model again; a store holding what is called memory; a task decomposition scheme, typically a prompt template instructing the model to produce a list of steps together with code that iterates over the list; termination conditions and budget limits; and an authorization scope determining which credentials the tools carry. This assembly is commonly called a *harness*, and the term is apt. It is a rig built around a component in order to direct and constrain it.

Two observations follow, and both matter for allocation.

The first concerns proportion. Of the code that determines what an agent does, the overwhelming majority is ordinary software: conditionals, loops, parsers, credential stores, retry logic, queue managers, error handlers. The model contributes a single function, invoked repeatedly: given a context, emit the next tokens. Everything that makes the assembly consequential, that it can send a message, transfer a sum, delete a record, or file a document, is a capability the harness conferred by wiring a credential to a tool. *The model did not acquire the ability to act in the world. The harness granted it*. Nothing in the weights differs between a model invoked once with no tools, which produces text, and the same model invoked in a loop with creden-

tialed tools, which produces consequences. The difference is entirely non-learned.

The second concerns the risk profile, which changes in kind rather than in degree. In a conversational deployment, a model error produces a false statement that a human may or may not act upon; the human is the actuator, and the harm passes through human judgment. In an agent deployment the harness is the actuator, and an error propagates to the world without a person in the path. Every available mitigation consequently lives in the harness: the scope of the credentials granted, whether a given class of action requires confirmation before execution, whether operations are reversible or destructive, whether spending and iteration are bounded, whether tools execute in a sandbox. *Not one of these is a property of the model, and the model's producer cannot know which of them a given harness author implemented*. The same weights sit behind an agent that drafts a document for review and an agent that executes payments unattended.

The development that makes this more than a technical curiosity is that building a harness no longer requires software engineering. Platforms now allow a person with no programming knowledge to assemble an agent by selecting a model, connecting tools through pre-built integrations, granting the credentials those integrations require, and describing the task in prose. In the terms of Section 2.9, the person who does this is the author of the application: they determine its purpose, the population it will affect, the scope of its authority, and whether any human confirmation stands between the model's output and the world.

Three consequences follow, and the third is the difficult one.

The chain lengthens by a party. Between the model's producer and the persons affected there now stand a platform and an agent author, and the platform is a component supplier to the agent author in precisely the way the model producer is a component supplier to the platform. The structure recurses.

The platform cannot know what will be built on it, and for the same structural reason established throughout this section: purpose is determined above. A platform offering a calendar integration and a messaging integration cannot infer whether the resulting agent will confirm dental appointments or send unsolicited commercial messages to a purchased list. Its epistemic position with respect to the agent author is the model producer's position with respect to the platform.

And the party who does determine purpose may have no capacity whatsoever to assess what it has built. This is genuinely new. Throughout Sections 2.2 to 2.9 the integrator was tacitly a professional software organization: a bank, a utility, a vendor, a public authority, with engineers and, at least in principle, the means to read a technical specification. The author of an agent assembled on a no-code platform may be a sole trader, a clinic receptionist, or a municipal employee, with no notion of what a false positive rate is and no means of evaluating a performance envelope even if one were placed in their hands. The framework developed in this chapter allocates the duty of assessing fitness to the party that determines purpose. Here that party determines purpose and cannot assess anything. The framework must answer for that, and Section 5.6 does.

2.11 Stage nine: deployment topology, and why the interface dominates

The same adapted model can be operated in at least four configurations: on hardware the deployer owns and administers; on cloud infrastructure the deployer rents but controls; through a third party that hosts open-weight models as a service; or through an interface operated by the model's own producer. These are ordinarily treated as commercial arrangements. They are in fact the stage at which observability, auditability, and the practical possibility of intervention are determined, and they are the subject of Section 4.2.

Why do hosted interfaces dominate? The answer is hardware economics, and it is worth being explicit because the legal literature tends to treat the prevalence of interfaces as a matter of convenience. A seventy-billion-parameter model at half precision requires on the order of one hundred and forty gigabytes of accelerator memory before any allowance for the caches that serving requires. A single tenant provisioning that capacity pays for it continuously and uses it intermittently. A provider aggregating thousands of concurrent requests across a farm of accelerators achieves utilization no individual tenant can approach, and the cost per unit of output falls accordingly. The interface is not primarily a product decision; it is a consequence of the cost structure of the underlying hardware.

The consequence for the integrator is that what it consumes is not a model but a

pipeline, and the pipeline is not observable. Behind an endpoint, the integrator cannot establish which version of the weights served a request, whether a router directed it to a different model entirely, whether retrieved context was injected before the prompt, whether a cache returned a stored response, whether input or output filters modified the text, or what system prompt was prepended. This is not speculation about bad faith; it is the ordinary architecture of a commercial serving stack.

It also has a methodological corollary worth recording, because it recurs in Section 4.2. The research literature on memorization is conducted almost exclusively on open-weight models executed locally, and for a reason of principle: a training sequence recovered through a proprietary endpoint cannot, even in principle, be distinguished from a sequence retrieved by a component sitting in front of the model. The opacity that matters here is not the opacity of a neural network. It is the opacity of a commercially confidential software pipeline. Legal discourse routinely treats these as one phenomenon, and they are two, with different causes and different remedies.

2.12 The cost gradient and its structural consequence

One quantitative feature cuts across every stage described above and deserves separate statement, because it explains the shape of the market that regulation must address.

Costs at these stages differ not by percentages but by orders of magnitude, and in two directions at once. *Cifra*'s model trains in minutes and answers in microseconds on ordinary processors. *Atendia*'s base model consumed millions of accelerator-hours to train. Within the generative pipeline, training is far costlier than inference (a single request against a served model costs a small fraction of a cent, against a training run measured in tens of millions of dollars), yet aggregate inference over a successful model's service life can exceed the cost of producing it. And inference on such a model remains three to four orders of magnitude more expensive per decision than inference on a logistic regression, which is precisely why classical methods continue to make most consequential automated decisions in finance and insurance, and why they deserve more regulatory attention than they currently receive.

The structural consequence is this. The cost gradient produces the population of

actors at each stage. Pre-training is available to a handful of firms with access to capital and hardware. Adaptation is available to anyone, and is therefore performed constantly, by parties no register records. Integration is performed by tens of thousands of ordinary software teams with no machine learning expertise at all. *The asymmetry of cost is an asymmetry of capability, and it is a reason of principle rather than convenience for allocating obligations by stage*. A duty specified once, for a single undifferentiated "provider," will either be unbearable for the integrator or trivial for the producer, and in practice tends to be both simultaneously.

Section 3 turns from what the stages are to what each of them can and cannot determine.

3. CONTROL, COMPOSITION, AND WHERE RISK BECOMES KNOWABLE

3.1 Three capacities, and why they come apart

Section 2 established what the stages are. Allocating duties requires a different question: not what each stage *is*, but what each stage *can determine*. Answering it requires distinguishing three capacities that legal analysis habitually treats as one.

Epistemic access is the capacity to know a risk-relevant property of the system. *Nordica*'s vendor knows the demographic composition of its corpus and its measured error rates by subgroup. The integrator who licenses it does not, and cannot acquire that knowledge by inspection of the artifact.

Causal access, or effective control, is the practical capacity to alter that property or its consequences. Once the model is released, the vendor cannot change its corpus; doing so would produce a different model. The integrator, by contrast, can change the matching threshold in an afternoon, and can change what happens on a match in a week.

Temporal position is the moment at which a decision closes. Corpus composition closes at collection. Capability closes at the end of the training run. Behavior never fully closes, because decoding parameters, prompts, thresholds, and filters remain mutable in production for the life of the deployment.

The central finding of this section is that these three capacities are *systematically distributed across different actors*, and that the law's habit is to impose the obligation on whichever actor is commercially most visible. Take a demographic performance disparity in a facial recognition deployment. Epistemic access to the disparity sits with the model producer, which measured it. Causal access over whether the disparity becomes a harm sits with the integrator, which sets the threshold, composes the gallery, and determines the consequence of a match. And the temporal position of the property is such that it was fixed before the integrator was even a party. No single actor can discharge a duty to "ensure the system does not discriminate," because discharging it requires all three capacities and no one holds all three.

This is a variant of the "many hands" problem, and the algorithmic accountability literature has described the difficulty with care, particularly in analyzing how contractual intermediation disperses decision authority across firms.[21] The diagnosis offered here differs in one respect that matters for the remedy. That literature decomposes the chain along commercial and contractual lines. The claim advanced here is that the load-bearing decomposition is *technical*: two firms occupying identical contractual positions may sit in entirely different positions in the control structure, depending on which stages of Section 2 each of them actually performs. A firm that licenses weights and applies an adapter is in a different control position from a firm that licenses an interface, even where both are "deployers" under identical terms. Contract is a poor proxy for control, and allocation by contract will therefore misfire even where the value chain is fully mapped.

3.2 The determination gradient: capability upward, purpose downward

The stages of Section 2 are not merely sequential; they determine different things,

---

[21] Jennifer Cobbe, Michael Veale & Jatinder Singh, *Understanding Accountability in Algorithmic Supply Chains*, in PROCEEDINGS OF THE 2023 ACM CONFERENCE ON FAIRNESS, ACCOUNTABILITY, AND TRANSPARENCY 1186 (2023). *See also* Jennifer Cobbe, Michelle Seng Ah Lee & Jatinder Singh, *Reviewable Automated Decision-Making: A Framework for Accountable Algorithmic Systems*, in PROCEEDINGS OF THE 2021 ACM CONFERENCE ON FAIRNESS, ACCOUNTABILITY, AND TRANSPARENCY 598 (2021).

and they do so in opposite directions.

Corpus, architecture, training, and post-training each successively narrow what the artifact *can* do. They constitute a capability. Nothing at any of these stages specifies what the capability is *for*. A facial embedding model does not know whether it will verify cadets or identify suspects. A language model does not know whether it will draft marketing copy or refuse to disclose a customer's arrears.

Application, interface, and deployment topology each successively specify what the capability will be used *for*: on whom, with what authority, with what consequence attaching to error, subject to what oversight and what remedy. Nothing at any of these stages alters what the artifact *can* do.

Risk requires both. It requires a capability with a failure mode, and a deployment in which that failure mode has a consequence. The relation is multiplicative rather than additive: if either factor is absent, the product is zero. A model with severe demographic performance disparities sitting unused on a storage volume harms nobody. A model of impeccable accuracy deployed to a purpose that is itself wrongful is not redeemed by its accuracy.

The two factors are determined by different actors, at different times, and neither actor has sight of the other's factor. That is the structural fact from which the whole of Section 5 follows.

Atendia illustrates it economically. The base model's producer determined that the artifact can generate fluent Portuguese prose about tariffs, and also that it will occasionally assert things that are false with the same fluency it asserts things that are true. The utility determined that the artifact's output would be presented to a customer as an authoritative statement about that customer's account. Neither the capability alone nor the presentation alone constitutes the risk. The risk is the pair, and the pair was never assembled in the mind of any single party.

3.3 Reversibility as an allocative criterion

A second criterion follows from the cost gradient described in Section 2.12, and it is almost entirely absent from current instruments.

Decisions at some stages are cheap to reverse. A decoding parameter can be

changed in seconds. A system prompt takes minutes. A matching threshold takes an afternoon, allowing for retesting. A guardrail or output filter takes days. An adapter takes hours to days, and can simply be detached.

Decisions at other stages are expensive to reverse or cannot be reversed at all. Corpus composition cannot be altered except by retraining. Architecture cannot be altered except by retraining. The training run itself is a capital event measured in months and in millions of units of currency, and its output is not amendable.

The regulatory consequence deserves blunt statement. *At reversible stages, the law can meaningfully require monitoring, correction, and iteration. At irreversible stages, it can require only ex ante diligence and accurate disclosure*. A provision that demands corrective action from a pre-training producer after release is either demanding the production of a different artifact or demanding nothing at all. This is not a loophole created by clever drafting; it is a physical property of the process. The AI Act's post-market monitoring and corrective-action architecture is well designed for the reversible stages and largely inapplicable to the irreversible ones, and the instrument does not distinguish between them.

The converse is more important, and it anticipates an objection to Section 5. Because irreversible stages cannot be corrected after the fact, *accurate disclosure is not a weak substitute for a real obligation at those stages; it is the only obligation that can be discharged there*. The choice at the pre-training layer is not between specification and something stronger. It is between specification and nothing.

3.4 Orthogonality: AI-intensity is not risk-intensity

The Introduction asserted that the two axes are independent. The assertion can now be made precise.

Call *AI-intensity* the proportion of a system's decision-relevant logic that is performed by learned components. Call risk-intensity a composite assessment reflecting the severity of the harm a failure can produce, the probability of that failure, the size and vulnerability of the affected population, the reversibility of the harm, and the availability of an effective remedy. Nothing in the definition of either quantity constrains the other.

The three worked systems occupy three different positions, and a fourth case completes the picture.

Cifra has very high AI-intensity: the entire substantive task, reading a handwritten numeral, is performed by a learned component, and there is no rule-based alternative in the pipeline. Its risk-intensity is negligible. Individual errors are small in magnitude, they are detected by ordinary reconciliation, they are reversible by a correcting entry, and they are remedied within a business day. Under every definition surveyed in Section 1, *Cifra* is an AI system, and it therefore passes the gate through which regulatory attention is directed.

Nordica installed at a Swedish military academy has high AI-intensity and negligible risk-intensity, for the reasons Section 4.1.2 sets out.

Nordica installed for street-level identification in Rio de Janeiro has *identical* AI-intensity, the same weights performing the same computation, and extreme risk-intensity. A false match against a police watchlist can produce detention, of a population that is large, involuntary, disproportionately exposed, and without recourse of the kind a correcting entry supplies. No property of the AI system distinguishes this case from the previous one, because the distinguishing properties are not properties of the AI system.

The fourth case runs the other diagonal. A fully deterministic, rule-based system that automates the calculation of a public benefit or a public debt, containing no learned component whatsoever, has AI-intensity of zero and therefore falls entirely outside the instruments considered in Section 1. It is nonetheless capable of producing harm at national scale, as the documented failures of automated public administration in several jurisdictions attest.[22] The regulatory apparatus constructed to govern algorithmic harm does not reach it, and the reason it does not reach it is that the apparatus selects on the wrong axis.

---

[22] Royal Commission into the Robodebt Scheme, Report (Commonwealth of Australia, July 2023). The scheme, introduced as an income-matching measure in 2015, derived a fortnightly income figure by averaging annual taxation data and raised a debt wherever that figure diverged from the amount a recipient had declared, placing on the recipient the onus of disproving the calculated debt. It contained no learned component. Debts of approximately AUD 1.76 billion were asserted against some 433,000 people before the scheme was abandoned.

The two failures are one failure seen from opposite sides. The deterministic system never entered scope; the learned system may leave it whenever leaving is cheaper than complying. Section 1 gave the mechanism: a fitted component swapped for an authored one, at some cost in accuracy and none in effect on the persons subject to the decision. This is regulatory arbitrage of an unusually clean kind, since it requires no jurisdictional relocation, no corporate restructuring, and no concealment. It requires only that the regulated party do slightly worse engineering, in public, for a stated reason. *A boundary that rewards a firm for degrading its own model is not merely misplaced; it is inverted*.

It may be objected that risk tiering corrects for this downstream, since a system inside the definition but outside the risk annexes bears little. The objection concedes the point: where the definitional gate does no allocative work, a filter keyed to use must do all of it. The conclusion is not that the definitions are imprecise. It is that **they sort systems along AI-intensity while purporting to govern them according to risk-intensity, over-including the innocuous, excluding the deterministic, and failing entirely to distinguish the two* Nordica *deployments**, which is the distinction that matters most.

3.5 Composition, and why component-wise assessment fails

The preceding subsections concern the distribution of control across stages. This one concerns something stronger: that risk is not a property of stages at all, but of their assembly.

Risk does not aggregate additively over components. Two components each safe in isolation can compose into a dangerous system, and two components each individually alarming can compose into a safe one. This is not a novel observation. It is the founding insight of system safety engineering, a discipline that abandoned component-reliability approaches to safety decades ago. Perrow argued that in tightly coupled, interactively complex systems, accidents are emergent structural properties rather than component failures.[23] Leveson developed the point into a general critique: treating safety as the sum of component reliabilities is a category error, because

[23] Charles Perrow, Normal Accidents: Living With High-Risk Technologies (1984).

safety is a control problem defined at the level of the system and its constraints.[24]

Cifra demonstrates the mechanism in the simplest possible form. The convolutional network is safe. The automatic posting routine is safe. The composition is safe only because a confidence threshold routes uncertain items to a human clerk. Change that one number (a numeric literal in non-learned code, requiring no modification to any model, no retraining, no new corpus, and no conformity reassessment of any component) and the system becomes capable of misposting deposits at scale. Nothing about any component changed. The risk changed entirely.

Nordica demonstrates the reverse. Its error profile becomes a discrimination harm only through a specific composition: a watchlist populated in a particular way, a threshold set at a particular point, an action that follows automatically on a match, and the absence of an independent corroboration step before that action. Break any one of those links and the harm does not occur. The error profile is unchanged in every case. What varies is the assembly.

Two consequences follow.

The first is that *component-wise conformity assessment cannot establish the safety of an assembly*. A regime that certifies models against one set of requirements and applications against another, and infers the safety of the composition, is committing precisely the error that aviation and process industries abandoned in the 1980s. Related concerns have been raised in the literature on sociotechnical fairness, which identifies the assumption that a technical solution validated in one social context will remain valid when transported to another as a distinct and recurring failure of framing.[25]

The second is that the human oversight on which so much of the regulatory architecture depends is itself a composed property, and a fragile one. Whether a clerk meaningfully reviews what is routed to her depends on the interface variables listed in Section 2.9, and the tendency of operators to defer to automated recommendations, increasing with the apparent reliability of the system, is among the better-

---

[24] Nancy g. Leveson, engineering a safer world: systems thinking applied to Safety (2011).

[25] Andrew D. Selbst et al., *Fairness and Abstraction in Sociotechnical Systems*, in PROCEEDINGS OF THE CONFERENCE ON FAIRNESS, ACCOUNTABILITY, AND TRANSPARENCY 59 (2019).

documented findings in human factors research.[26] The oversight the law assumes is therefore designed at a stage where no learned component is involved at all.

3.6 The limit of the thesis: capability-intrinsic risk

The proposition that risk becomes knowable only at the application layer is not universally true, and an argument that fails to say so invites a decisive objection. The exception must be stated, and stating it refines the thesis rather than weakening it.

Some risks attach to a capability itself, independently of any application. A model that materially assists in the synthesis of a dangerous pathogen, or that can autonomously discover and exploit software vulnerabilities at scale, carries a risk that does not depend on which application invokes it, because the harm requires no particular deployment design. Access to the capability is sufficient. For risks of this class there is no downstream integrator whose choices could render the capability safe, and consequently the producer both can and must assess: it alone holds the artifact before distribution, and there is no later stage at which the assessment could be performed.

Two classes of risk must therefore be distinguished, and they attract different duties.

Context-dependent risks (accuracy, demographic performance disparity, adequacy of oversight, proportionality of purpose, availability of remedy) are constituted by the relation between a capability and a deployment. They are not knowable upstream, and the producer's duty in respect of them is a duty of specification.

Capability-intrinsic risks, meaning dual-use capabilities whose harm potential is realized by access rather than by deployment design, are knowable upstream. The producer's duty in respect of them is a duty of evaluation, and where appropriate of restriction on release.

It is worth noting, to the advantage of the argument, that the AI Act's regime for general-purpose models presenting systemic risk is the one place in the instrument where obligations are keyed to *capability* rather than to commercial role, and it is the

[26] Raja Parasuraman & Victor Riley, *Humans and Automation: Use, Misuse, Disuse, Abuse*, 39 HUMAN FACTORS 230 (1997).

place where the instrument's structure is most defensible.[27] The difficulty is that the same statute reverts to role-based allocation for everything else. The framework proposed here is not a repudiation of the AI Act's architecture but a generalization of the part of it that already works.

The thesis in its final form is therefore this: *risk is knowable only at the application layer, save for that subclass of risks realized by access to a capability rather than by the design of its deployment*.

3.7 The allocative criterion, and what follows from it

The material of this section can be reduced to a single test.

An obligation is well allocated when it is imposed on an actor who holds, at a moment when action remains possible, both epistemic access to the relevant property and causal access to alter that property or its consequences.

Applied to current instruments, the test is failed routinely, and it is failed in a specific and diagnosable way: knowledge and control sit with different parties. That is not an accident of drafting but the ordinary condition of a layered technical process in which capability is produced by one set of actors and aimed by another.

Where knowledge and control are separated, the law has two options. It can impose the duty on one of the parties and accept that it will not be discharged, which is in substance the present approach, and which generates documentation that satisfies auditors without affecting outcomes. Or it can impose duties whose function is to *reunite* knowledge with control: a duty on the knowing party to transmit what it knows, in a standardized and comparable form that the controlling party can act upon, and a correlative duty on the controlling party to act upon it.

This reframing is the point of the proposal in Section 5, and it can be stated in a sentence. *The function of a specification duty is not to allocate blame but to move knowledge to where control already sits*. A selection duty is its necessary correlate: information transmitted to a party under no obligation to use it changes nothing.

Before turning to that proposal, the framework must be tested. Section 4 does so

---

[27] Regulation (EU) 2024/1689, *supra* note 1, arts. 51-55.

through two cases chosen because each isolates one axis of the analysis while holding the other constant: the axis of purpose, and the axis of control.

4. TWO ILLUSTRATIONS

The framework of Section 3 makes a testable claim: that risk is constituted by the relation between a capability and a deployment, and that the parties holding knowledge of each are different. Two cases test it, and they are constructed as controlled comparisons. The first holds the artifact constant and varies the purpose. The second holds both the artifact and the purpose constant and varies only where the computation physically occurs. If the framework is correct, risk should vary substantially in both cases while the object that current instruments regulate remains unchanged.

4.1 Facial recognition: the axis of purpose

4.1.1 What a facial recognition system actually is

Public and legislative discussion treats facial recognition as one technology. It is at least three, distinguished by the question asked of the model, and their error characteristics differ so widely that a performance figure valid for one is misleading for another.

Verification compares a presented face against a single enrolled reference and answers whether they are the same person. This is the task performed when a telephone is unlocked. The comparison is one-to-one.

Closed-set identification compares a presented face against a gallery of enrolled identities and answers which of them it is, on the assumption that it is one of them. This is the task performed at a controlled entrance with a known roster. The comparison is one-to-many, but the correct answer always exists in the gallery.

Open-set identification, or watchlist search, compares a presented face against a gallery and answers whether it is any of them, where the correct answer, for almost every face presented, is *none*. This is the task performed in public-space surveillance, and it is qualitatively harder than either of the others.

The reason is arithmetic and deserves setting out, because it is the single most consequential technical fact about the deployment discussed below. If each individual comparison carries some probability of a false match, then searching one face against a gallery of many identities carries a probability of at least one false match that grows with the size of the gallery. A per-comparison false match rate low enough to be irrelevant in one-to-one verification becomes a routine occurrence when the same threshold is applied against a gallery of tens of thousands. Standardized evaluation programs accordingly measure one-to-one verification and one-to-many identification as distinct tasks with distinct metrics.[28]

Two further properties of these systems matter here. First, demographic performance differentials in facial recognition are measurable and have been measured at scale: a large evaluation of commercial and academic algorithms documented substantial variation in false match rates across demographic groups, with the magnitude of the differential itself varying enormously between algorithms, some exhibiting pronounced disparities and others none detectable.[29] That last observation is the important one for this chapter's purposes. A demographic differential is not an inherent feature of the technique. It is a measurable, reportable, and improvable property of a specific artifact, which is precisely what makes it a candidate for specification.

Second, as established in Section 2.3, the model itself returns neither a match nor an identity. It returns a vector. The threshold that converts a distance between vectors into a verdict is set outside the model, by the integrator.

---

[28] NAT'L INST. OF STANDARDS & TECH., Face Technology Evaluations: FRTE / FATE, https://www.nist.gov/programs-projects/face-technology-evaluations-frtefate. The program formerly designated the Face Recognition Vendor Test was renamed and divided in August 2023: tracks concerning identity verification run under FRTE, tracks concerning image analysis under FATE. FRTE reports 1:1 Verification and 1:N Identification as separate ongoing tracks with separate metrics, and maintains a distinct Demographic Effects track.

[29] Patrick Grother, Mei Ngan & Kayee Hanaoka, *Face Recognition Vendor Test (FRVT) Part 3: Demographic Effects*, NISTIR 8280 (Nat'l Inst. of Standards & Tech. 2019).

4.1.2 Deployments A and B: the benign cases

In the first deployment, a model performing one-to-one verification unlocks a telephone. The gallery has size one; enrollment is performed by the data subject, of the data subject, revocably; the threshold is tuned by the device manufacturer toward rejecting rather than accepting on uncertainty; a false acceptance presupposes physical possession of the device; and a fallback is always available, since a false rejection means the user types a passcode. Whatever demographic differential the model carries is present here too, manifesting as a modest difference in convenience: some users are asked for their passcode more often than others. That is worth naming rather than dismissing, but it is not a matter for which a conformity assessment regime is the proportionate response.

In the second, the same class of model performs closed-set identification against a roster of cadets at a Swedish military academy, and the model is the one the Introduction posited: trained on a corpus drawn predominantly from Northern European adult men. The gallery is small and known; enrollment is institutional but consensual, of a population contractually bound and individually identifiable; a false rejection means a cadet presents a card; and a false acceptance is caught by a physical checkpoint operating independently of the biometric system. Critically, the deployment population and the training population coincide almost exactly, so the model's performance envelope is well matched to its use, and its demographic differential is close to irrelevant, not because it has been mitigated but because the population it disadvantages is barely present. **The model has not become fair. The deployment has become fortunate.** The distinction is the crux of the argument and should not be softened: a model carrying a severe differential can be entirely appropriate for a particular use, and its appropriateness there tells one nothing whatsoever about its appropriateness elsewhere.

4.1.3 Deployment C: watchlist search in public space

The same weights are integrated into a public security system in Rio de Janeiro. Cameras in transit stations and public thoroughfares present faces continuously;

each is searched against a watchlist of persons subject to outstanding warrants. Deployments of this kind are not marginal in Brazil: monitoring projects record several hundred active systems with a potential reach of some two-fifths of the national population.[30]

Every risk-determining variable has changed, and none of them is a property of the model.

The *task* has changed, from verification or closed-set identification to open-set search, which means the performance figure the vendor measured is not the performance figure that governs.

The *population* has changed, from Northern European to Brazilian, which is among the most phenotypically heterogeneous populations in the world and bears no resemblance to the training distribution. The model's demographic differential, immaterial in Deployment B, now falls on the majority of the people it observes, and falls hardest on the groups already most exposed to police attention.

The *gallery* has changed, in size and in the manner of its construction. Enrollment is performed by the state, of third parties, without consent, from records whose accuracy and currency are outside the system's control.

The *base rate* has changed, and this is the variable that most sharply demonstrates the chapter's thesis. In a population of a hundred thousand daily passengers of whom a small number are on the watchlist, the prior probability that any given face is a watchlist member is very low. The probability that a given alert is correct, the positive predictive value, depends on that base rate, not merely on the false match rate. A system with an impressively low false match rate, applied to a population in which almost nobody is sought, will generate alerts that are predominantly false. *And the base rate is a property of the deployment: it depends on the observed population and the composition of the watchlist, neither of which the model's producer knows or could*

---

[30] Defensoria Pública da União & Centro de Estudos de Segurança e Cidadania, Mapeando a Vigilância Biométrica: Levantamento Nacional Sobre o Uso do Reconhecimento Facial na Segurança Pública (Rio de Janeiro, 7 May 2025), https://direitoshumanos.dpu.def.br/wp-content/uploads/2025/05/CESeC__DPU_Mapeando_a_vigilancia_biometrica.pdf, recording 376 active projects as of April 2025 with a potential reach of some 83 million people, approximately 41 percent of the national population.

know. It follows that even a perfectly specified model, whose false match rate has been measured and honestly disclosed, cannot tell an integrator the probability that any given alert is correct. That calculation requires a variable available only downstream.

The *consequence of error* has changed, from typing a passcode to being stopped, questioned, or detained. The harm is not reversible by a correcting entry, and effective remedy is uncertain. Brazil has operational deployments of this kind, and misidentifications with serious consequences have been documented.[31]

The *fallback* has disappeared, and in this jurisdiction so has the general data protection framework. There is no equivalent of the passcode: the person misidentified has no alternative channel through which to establish that the system is wrong before the consequence attaches. Nor is there a statutory one. Brazil's data protection statute expressly excludes from its scope processing carried out for the exclusive purposes of public security, national defense, state security, and the investigation and repression of criminal offenses, and provides that such processing is instead to be governed by specific legislation observing proportionality, due process, and the rights of the data subject. That specific legislation has not been enacted.[32] The deployment carrying the gravest consequences examined in this chapter is therefore not merely governed by an instrument that selects the wrong subject. It is governed by no data protection instrument at all.

And the *composition* determines everything, exactly as Section 3.5 predicted. A system that surfaces a ranked list of candidates to a trained analyst who must independently corroborate before any action, and a system that dispatches a patrol on the strength of a single top-ranked hit, may contain identical weights, identical thresholds, and identical error rates. They are not the same system in any sense the law

---

[31] Mapeando a Vigilância Biométrica, supra note 30, reporting that more than half of the police stops motivated by facial recognition surveyed resulted in mistaken identification, and that no state publishes figures on identification errors, false positives, or wrongful detentions. The underlying case series is maintained by O Panóptico, a project of the Centro de Estudos de Segurança e Cidadania monitoring the adoption of facial recognition by Brazilian public security institutions since 2019, which records 24 documented system failures between 2019 and April 2025. The most widely reported concerns João Antônio Trindade Bastos, removed from a stadium in Aracaju by military police in April 2024 on a false match, after which the state suspended the deployment.

[32] Lei No. 13.709/2018 (Lei Geral de Proteção de Dados Pessoais), art. 4, III(a) and § 1.

should care about, and the difference between them was authored entirely in application code. The standardized evaluation program treats these as distinct measurement scenarios for exactly this reason: one-to-many search is reported separately for a thresholded configuration, appropriate where the output will drive a largely automated decision, and for an unthresholded configuration returning a fixed number of candidates, appropriate where a human examiner is assumed to review every result. [33] The same algorithm appears under both, with different metrics, because the presence of human review changes what counts as an error. The measurement infrastructure has internalized a distinction that the allocation of legal duties has not.

4.1.4 What varied, and what did not

Across the three deployments, the following did not change: the architecture, the training corpus, the weights, the measured accuracy on the vendor's benchmark, the license, and the vendor's technical documentation.

The following changed: the task asked of the model, the deployment population, the construction of the gallery, the base rate of positives in the observed population, the threshold, the consequence of error, the reversibility of harm, the availability of a fallback, the existence of independent corroboration, and the availability of remedy.

Every variable in the second list is a determinant of risk. Not one of them is a property of the AI system as defined by the instruments surveyed in Section 1. The vendor could not have assessed any of them, because at the time it fixed the artifact, none of them existed.

The conclusion is not that the vendor bears no responsibility. It is that the vendor's responsibility cannot be a responsibility for risk, because risk was not yet determinate. It can only be a responsibility for the accuracy and completeness of what it says about the artifact, and the material of 4.1.1 tells us what that statement must contain if it is to be usable. It must specify the *task* to which the reported performance applies, because a verification figure does not govern an open-set search. It must

---

[33] Nat'l Inst. of Standards & Tech., Face Recognition Technology Evaluation (FRTE) 1:N Identification, supra note 28, which reports the one-to-many track separately for a thresholded configuration and for an unthresholded configuration returning a fixed number of candidates.

report performance *by demographic group*, because a single aggregate figure conceals the variable that determines whether the model is suitable for a given population. It must state the *threshold* at which each figure was obtained, because performance without a threshold is not a figure at all. And it must be produced by a *standardized method*, because a figure that cannot be compared against a competitor's is a marketing claim rather than a specification.

This is the tire's sidewall, and Section 5 develops it.

4.1.5 What the AI Act gets right, and where it stops

It should be acknowledged that the European instrument handles this case better than its general architecture would suggest. It prohibits certain uses of real-time remote biometric identification in publicly accessible spaces for law enforcement purposes, subject to exceptions, and it classifies remote biometric identification systems as high-risk.[34] That is regulation by *use*, and it is correct: the legislator evidently understood that the danger resides in the deployment rather than in the technique.

Yet having located the risk in the use, the instrument allocates the obligations by commercial role. The vendor is a provider and bears provider obligations; the police authority is a deployer and bears deployer obligations. What the deployer needs in order to evaluate its own deployment (a task-specific, demographically disaggregated, threshold-referenced, comparably measured performance envelope) is not what the provider is required to hand it in the form in which it is required to hand it. The instrument's diagnosis is right and its transmission mechanism is underspecified. This internal tension, rather than any external critique, is the strongest available argument for the reform proposed below.

---

[34] Regulation (EU) 2024/1689, *supra* note 1, art. 5(1)(h) and Annex III(1). The prohibition is in art. 5(1)(h), subject to the three exhaustively listed objectives in points (i)-(iii), and to the safeguards in art. 5(2)-(7), including necessity and proportionality, a fundamental-rights impact assessment, registration, and prior independent authorisation save for narrowly defined urgency.

4.2 One model, three topologies: the axis of control

The second comparison holds more constant than the first. Here the artifact, the purpose, the affected population, and the consequence of error are all fixed. Only the physical location of the computation varies.

Consider a law firm deploying an open-weight language model as a document review assistant over client files. The model chosen is an openly released set of weights of the kind published by DeepSeek. The case is chosen because such a release makes all three topologies genuinely available, which a proprietary model does not. [35]

4.2.1 Three configurations

On-premises. The firm downloads the weights, verifies their cryptographic hash, and executes them on accelerators it owns, within a network segment it administers.

Third-party hosting. The firm sends requests to an inference provider that hosts open-weight models as a service, executing the same nominal weights on the provider's hardware.

Original developer's interface. The firm sends requests to an endpoint operated by the entity that trained the model, executing on that entity's hardware.

4.2.2 What varies

Artifact identity. On-premises, the firm can pin an exact file, verify it by hash, and hold it unchanged for years. In the hosted configurations, the firm receives an assertion about which weights are serving its requests and has no means of verifying it, nor of detecting a silent update. A recurring practical consequence is that a prompt

[35] DeepSeek-R1 is released with both the code repository and the model weights under the MIT License, permitting commercial use, modification, and derivative works including distillation. Note that the R1-Distill variants are derived from other base models and carry their upstream licenses.

validated in testing degrades weeks later with no change on the firm's side.

Pipeline composition. On-premises, the firm authored or selected every component: it knows which runtime executes the weights, which chat template is applied, what sampling parameters are set, whether any retrieval component precedes the model, and what system prompt is prepended, because it wrote them. Behind either hosted endpoint, none of this is observable. A router may direct the request to a different model. A cache may return a stored response. Input and output filters may modify text. A system prompt may be injected. As Section 2.11 noted, this is the ordinary architecture of a commercial serving stack rather than an allegation of bad faith.

Data flow. On-premises, client material does not cross the firm's perimeter, which for a law firm is not a compliance preference but a professional obligation. In the hosted configurations it crosses, and the firm's protection is contractual rather than architectural.

Auditability. On-premises, the firm can reproduce any past output deterministically by fixing the weights, the parameters, and the random seed, which means an allegation about what the system did on a given date is testable. In the hosted configurations it cannot, because the pipeline that produced the output may no longer exist. The methodological point recorded in Section 2.11 becomes concrete here: on-premises, the firm can itself measure the model's propensity to reproduce training sequences; through a proprietary endpoint, an output that reproduces a copyrighted text cannot even in principle be attributed to memorization rather than to retrieval by an intervening component. The distinction is not academic where the consequence is an infringement claim.

Continuity and intervention. On-premises, the artifact persists and every parameter is modifiable. Through an endpoint, the model may be deprecated on notice, and the firm may adjust only what the interface exposes.

Jurisdiction. The three configurations may place the computation, and the client material, in three different legal orders.

One asymmetry cuts the other way, and honesty requires stating it. In the on-premises configuration the firm may run a four-bit quantized conversion of the weights on affordable hardware, which, as Section 2.8 established, is not the artifact

the developer released and tested. Local execution maximizes control and does not preserve artifact identity. Control and fidelity are distinct goods, and topology trades between them.

4.2.3 What the law sees

Under the taxonomy of the AI Act, the firm is a deployer in all three configurations, and its obligations are the same in all three. Its *capacity to discharge* them differs by orders of magnitude. Asked to produce records demonstrating how a particular output was generated, the on-premises firm can produce a complete and verifiable account; the firm using a proprietary endpoint cannot produce one even in principle, because the information does not exist on its side of the boundary. An obligation identical in text is satisfiable in one case and unsatisfiable in the other, and nothing in the instrument registers the difference.

A further consequence deserves separate statement, because it identifies a defect in the present situation rather than an inconsistency in its regulation. An integrator building on a frontier proprietary model cannot obtain the information that Section 4.1.4 identified as necessary, not because it negotiated badly but because no obligation to supply it exists and no commercial leverage can compel it. For the leading closed models, the architecture is described only in outline, the parameter count is unpublished, the composition of the training corpus is undisclosed, the post-training procedure is proprietary, and performance is reported on benchmarks of the producer's own selection rather than disaggregated by the dimensions a deployer would need.[36] A hospital, a bank, or a public authority building on such a model is in the position of a vehicle constructor forbidden to ask what the tire will withstand. And its position is worse than in either open-weight configuration, in a way that is easily missed: where weights are held, an integrator that was not told something can *measure* it, and independent evaluation substitutes imperfectly but genuinely for disclo-

[36] OpenAI, GPT-4 Technical Report, arXiv:2303.08774 (2023). The report states that, given the competitive landscape and the safety implications of large-scale models, it contains no further details about the architecture, including model size, nor about hardware, training compute, dataset construction, or training method.

sure. Behind a proprietary endpoint neither route is open. *The closed interface is the only configuration in which epistemic access is foreclosed by non-disclosure and by non-inspectability at the same time*, which is why the case for a mandatory specification duty is strongest precisely where the industry's resistance to it is strongest, and why Section 5.5 must confront the objection that such a duty would trespass upon commercial secrecy.

The reverse asymmetry is more instructive still, because it reveals what is wrong with the shape of the open-source exemptions rather than merely with their scope. In the on-premises configuration the original developer has no relationship with the deployment whatsoever. It does not know the firm exists. It cannot revoke access, observe misuse, or deliver a correction. Its causal access has fallen to zero. The intuition embodied in the open-source exemptions, that the developer of freely released weights should not bear the obligations of a provider placing a product on the market, correctly registers this.[37]

But the intuition is only half right, and the half it gets wrong is the operative one. *Open release eliminates what the developer can do; it does not diminish what the developer knows*. The developer retains sole and complete epistemic access to the corpus, the training procedure, the measured performance envelope, and the known failure modes. It is the only party in the world in a position to state those things, and its inability to control downstream use is precisely the reason its duty should be a duty to state them. The exemptions, as drafted, relieve the developer of disclosure at the moment when disclosure has become the only thing it is capable of providing. The correct move is not to exempt but to *convert*: from a duty of control, which release has made impossible, into a duty of specification, which release has made indispensable.

4.2.4 Why the least auditable configuration prevails

A final observation completes the argument. Of the three configurations, the one offering the weakest observability, the weakest auditability, and the weakest guaran-

[37] Regulation (EU) 2024/1689, *supra* note 1, art. 2(12); *see also id.* art. 53(2) and art. 25 (on modification and the assumption of provider obligations).

tee of artifact identity is the one most widely adopted and, for most organizations, the only economically available option. The reason was given in Section 2.11: a large model requires accelerator memory that a single tenant must provision continuously and will use intermittently, while a provider aggregating many tenants achieves utilization no individual tenant can approach.

The cost structure of the hardware therefore selects for the topology with the least observability, and no current instrument prices that externality. Auditability is treated as a compliance obligation of the deployer while being, in practice, a function of a procurement decision that economics has largely made in advance. A regime serious about auditability would have to attach consequences to topology, for instance by requiring that an integrator unable to observe its own pipeline obtain, by contract or by regulation, an attested description of it. That such a requirement sounds unusual is a measure of how thoroughly the layer has escaped legal attention, not of how difficult it would be to impose.

Section 5 turns from diagnosis to proposal.

5. THE DUTY OF SPECIFICATION AND THE DUTY OF SELECTION

5.1 An existing solution to the same problem

The difficulty identified in Sections 3 and 4 is not new, and the law has solved it before. Whenever a complex product is assembled from components supplied by parties who do not know the product's final application, the same question arises: on what terms does a component supplier answer for harm caused by the assembled whole?

The settled answer, in the products liability tradition, is that the supplier of a component is not liable for defects of the integrated product merely because its component was incorporated. Liability attaches where the component was itself defective, or where the supplier substantially participated in the design of the integra-

tion such that it can fairly be treated as a designer of the final product.[38] The doctrine reflects an allocation of exactly the kind Section 3 recommends: the component supplier is held to what it can know and control, the properties of its own artifact, while the integrator, who selected the component, specified the application, and knew the conditions of use, answers for the fitness of the assembly.

The proposal advanced here is that models be treated as components under this logic, and that the correlative duties be given determinate content. It is worth being explicit that this is an argument from continuity rather than from exception. The claim is not that artificial intelligence requires a novel liability regime; it is that a doctrine already available has not been applied, because the definitional apparatus surveyed in Section 1 constructed a single "AI system" with a single provider and thereby concealed the component relationship the doctrine presupposes.

5.2 Why the speed rating works

The proposal's shape is best conveyed by an instrument that has governed a genuinely analogous relationship for decades without controversy.

A tire manufacturer does not assess whether a given vehicle will be driven on an unrestricted autobahn, whether its owner drives fast, or what the consequence of a failure would be. It molds a speed rating into the sidewall. A vehicle constructor designing a car capable of 250 km/h selects a tire rated for at least that speed, and if it fits one rated for 210 the resulting liability is the constructor's, not the manufacturer's. The manufacturer discharged its duty by rating correctly.

What makes this arrangement work is not the existence of a specification. Four properties do, and each has a direct analogue in what is proposed below.

---

[38] RESTATEMENT (THIRD) OF TORTS: PRODUCTS LIABILITY, *supra* note 7, § 5. Brazilian law contains no direct analogue to the Restatement's component-parts rule. Lei No. 8.078/1990 imposes strict liability on manufacturers, producers, constructors and importers for product defects (art. 12), assigns merchant liability in specified cases (art. 13), and provides solidarity among multiple authors of consumer harm (arts. 7, sole paragraph, and 25, § 1). The treatment of a component supplier therefore turns on whether its own product was defective and causally connected to the harm, rather than on an express statutory component doctrine.

The rating is *standardized*. It is not a claim each manufacturer frames as it prefers; it is determined by a defined test method under a shared regulatory instrument.[39] A rating from one manufacturer means the same thing as a rating from another.

It is *comparable*. Because the method is common, a constructor can rank alternatives on the dimension that matters. A specification that cannot be compared is not a specification; it is advertising.

It is *mandatory*. It is not supplied where favorable and omitted where awkward. It is present on every tire sold.

And it is *legally consequential in itself*. A false or absent rating is a violation independent of any accident. The duty is not to avoid harm; it is to state the envelope accurately, and the breach is complete at the misstatement.

Note what the rating does not do. It does not reveal the rubber compound, the belt construction, the curing schedule, or anything else a manufacturer would regard as proprietary. It states what the artifact will withstand and is silent as to how that was achieved. *Performance specification is orthogonal to process secrecy*, a point to which Section 5.5 returns because it disposes of the objection most likely to be raised against the proposal.

5.3 The content of a model specification

Section 4.1.4 derived the required content from the facial recognition case rather than postulating it. Generalized, a model specification adequate to support downstream selection would state at least the following.

The *task* to which each reported figure applies. A one-to-one verification figure does not govern an open-set watchlist search, and a benchmark score on curated question-answering does not govern extraction of clauses from contracts. This is the single most frequently violated requirement in current practice, and the omission is not innocuous: it permits a figure obtained on an easy task to be relied upon for a hard one.

The *threshold or operating point* at which each figure was obtained, since a false

[39] U.N. Econ. Comm'n for Eur., Regulation No. 30, *supra* note 6.

match rate reported without a threshold is not a quantity.

Performance *disaggregated* along the dimensions on which it is known to vary: demographic group where the artifact processes human attributes, language and register where it processes text, and the operating conditions relevant to the modality. A single aggregate figure conceals precisely the variable that determines suitability for a given population.

The *training distribution* in terms sufficient to identify the population the artifact represents. This need not, and in most cases should not, be a disclosure of the corpus itself. It is a characterization: what population, what period, what languages, what provenance categories, what consent or licensing basis.

Known *failure modes and out-of-envelope conditions*: the conditions under which the artifact is known to degrade, and those under which its behavior has not been characterized at all. The second is as important as the first, and its absence is what makes current documentation misleading rather than merely incomplete.

Measured *memorization propensity*, on the reasoning of Section 2.6: the legally tractable property is not whether the artifact "contains" training data but its quantifiable tendency to reproduce it, which is testable and reportable.

Artifact identity sufficient to establish what was actually evaluated: a cryptographic digest, the serialization format, and the numerical precision. Section 2.8 established that a quantized conversion is not the artifact that was tested; a specification that does not identify its own referent cannot be relied upon.

And, where the model is served through an interface rather than delivered, an *attested description of the serving pipeline*: whether retrieval is interposed, whether outputs are filtered, whether a system prompt is prepended, whether requests may be routed to a different model, and the policy governing changes to any of these. Section 4.2 showed that without this the integrator's own obligations are unsatisfiable in principle.

5.4 Why current model cards are not specifications

The practice of publishing model documentation is well established, and the

proposal that originated it was sound.[40] But measured against the four properties of Section 5.2, current documentation fails on all four.

It is *not standardized*: there is no common test method, so two producers reporting "accuracy" may be reporting incommensurable quantities. It is *not comparable*: benchmarks are selected by the producer, which means the choice of what to measure is itself a discretionary disclosure. It is *not mandatory*: it is published where reputationally useful and thinned where not. And it is *not independently consequential*: an omission or an over-claim in a model card gives rise, at present, to no liability distinct from whatever harm may eventually follow.

The final defect is the structural one, and it explains the other three rather than merely joining them. *Where disclosure is voluntary, what is disclosed is negatively correlated with its regulatory value*. A producer publishes the favorable benchmark and omits the unfavorable demographic differential, not from bad faith but because the incentive operates exactly as one would predict. It follows that the cases in which specification matters most are precisely the cases in which voluntary disclosure fails. No amount of good practice, industry self-organization, or reputational pressure corrects an incentive that points the wrong way; only an obligation does.

5.5 Trade secrecy, enforcement, and why disclosure must be mandatory

The principal objection to a mandatory specification duty is that it would compel disclosure of commercially sensitive information. The objection is answered not by balancing but by the *form* of the duty, and the tire carries the answer precisely.

The speed rating discloses nothing about how the tire is made. It discloses what the tire will withstand. A model specification of the kind set out in Section 5.3 is constructed on the same principle: it reports measured envelope, not method. It requires the producer to state that performance degrades on a given population, not

[40] Margaret Mitchell et al., *Model Cards for Model Reporting*, in PROCEEDINGS OF THE CONFERENCE ON FAIRNESS, ACCOUNTABILITY, AND TRANSPARENCY 220 (2019); *see also* Timnit Gebru et al., *Datasheets for Datasets*, 64 COMMUNICATIONS OF THE ACM 86 (2021).

to reveal the corpus; to state the measured memorization rate, not to publish the training data; to state that a serving pipeline interposes retrieval, not to disclose the retrieval index. Everything a producer would genuinely wish to protect (architecture in detail, data acquisition arrangements, training recipe, optimization techniques) remains protected. What ceases to be protected is the *concealment of limitations*, which is not a trade secret in any defensible sense but simply an information advantage over the counterparty.

There is a second reason the duty must be mandatory rather than contractual, and Section 4.2.3 established it: the integrator has no leverage to obtain this information by negotiation. For the leading closed models it cannot be told and cannot find out, since non-disclosure and non-inspectability foreclose both routes at once. A regime that leaves the question to contract therefore leaves it to the party with all the bargaining power, and the predictable equilibrium is the one now observed, in which a hospital or a public authority deploys a component whose limitations it is structurally unable to learn.

Three enforcement features follow from the structure of the duty rather than from any general preference for stringency.

The breach is complete at the misstatement. Because the duty is to state accurately, it is violated when the statement is false or incomplete, independently of whether harm follows. This matters because harm at the model layer is mediated by downstream choices, so a harm-contingent duty would almost never be enforceable against a producer, and this is exactly why the present regime yields documentation without consequence.

Testability is a condition of the duty, not an addition to it. A specification no one can verify is an assertion, and Sections 2.11 and 4.2.2 established why verification is foreclosed behind a proprietary endpoint. A specification duty must therefore be accompanied by a means of verification proportionate to the risk class: for high-risk applications, evaluator access to the artifact or to a test environment with a pinned version and a declared pipeline. This is the point at which the AI Act's provisions on regulatory sandboxes and researcher access become structurally relevant, though they are currently framed around the model rather than around the serving pipeline

where the indeterminacy actually resides.[41]

The duty scales with capability, not with corporate form. This is developed in Section 5.7.

5.6 The correlative duty of selection

A specification duty on its own accomplishes nothing. Information transmitted to a party under no obligation to act upon it changes neither behavior nor outcome, and the history of mandated disclosure in consumer and financial regulation is largely a history of documents that were supplied and not read.

The correlative duty is one of selection, and it is the vehicle constructor's duty. An integrator deploying a learned component would be required to establish, and to be able to demonstrate, that the component's specified envelope covers the conditions of its intended deployment: that the population it will encounter falls within the population the artifact represents; that the task it will perform is the task for which performance was reported; that the operating point selected is the one at which the relevant figures were measured; and that, where the specified performance does not cover the deployment, either the deployment is altered or a different component is selected.

Three features of this duty deserve emphasis.

It is a duty of *fit*, not of outcome. The integrator is not required to guarantee that no error occurs. It is required to have selected a component whose stated envelope covers its use, and to have designed the assembly so that errors within the envelope do not produce disproportionate harm. That is the constructor's obligation when it selects a tire, and it is discharged by reasoning correctly from a specification, not by achieving perfection.

It requires *inputs the integrator alone holds*. Section 4.1.3 showed that the positive predictive value of an alert depends on the base rate of positives in the observed population, a quantity available only downstream. The duty of selection is therefore

[41] Regulation (EU) 2024/1689, *supra* note 1, arts. 57-61 (regulatory sandboxes and testing in real-world conditions).

not a duty to read a document; it is a duty to combine the producer's envelope with the deployer's own knowledge of population, base rate, consequence, and remedy. Neither party could perform this calculation alone, which is the whole point.

And it *attaches to the composition*, consistently with Section 3.5. The integrator answers for the assembly: the threshold, the action on output, the presence or absence of independent corroboration, and whether the interface supports the oversight the design assumes.

There is, however, a boundary condition, and Section 2.10 exposed it. The duty of selection presupposes a party competent to perform it. On an agent platform, the person who determines purpose, grants credentials, and decides whether any action requires confirmation may be a sole trader or a municipal employee with no means of reading a performance envelope, let alone of reasoning from one. A duty to select correctly cannot be discharged by a party unable to evaluate the specification, and imposing it anyway would generate precisely the empty compliance that Section 5.4 diagnosed in voluntary model cards, a form completed because a form was required.

Where the presumption of competence fails, the appropriate response is not to disclose more insistently but to change the character of the upstream duty: *to constrain the envelope rather than merely describe it*. The platform would be required to restrict what can be assembled (which tools may be combined, which credentials may be granted without independent verification, which classes of action may execute without human confirmation, what spending and iteration limits apply by default) rather than to inform a user who cannot act on the information. This is the ordinary structure of consumer product regulation, in which configurability is curtailed exactly where the configurer cannot be assumed to understand the consequences of configuring badly. It identifies the framework's own limit, and the limit is worth stating plainly: *specification is the right duty between competent parties; where competence is absent, constraint takes its place*.

5.7 The limits of the analogy

A tire has a single function and a measurable envelope. A general-purpose model has an open input space and capabilities its producer did not anticipate. The objection that such a model is therefore *unspecifiable* is the strongest available against this

proposal and must be met directly.

It is met by observing that the duty of specification scales with the specificity of the capability claimed, and that this scaling is a feature of the layered framework rather than a concession extracted from it.

Where a model is built and marketed for a defined task, such as facial verification, document classification, or credit scoring, the envelope is narrow and genuinely measurable. Standardized evaluation of exactly this kind already exists and has existed for years in the biometric domain, which demonstrates that the objection is not a technical impossibility but an absence of obligation.

Where a model is general in capability, the duty converts in content while remaining a duty of accurate statement. It becomes a duty to disclose *what was evaluated and what was not*: which benchmarks were run and on what versions, which failure modes are known, which populations and languages were assessed, which capabilities were tested for and which were not examined at all. The final element is the most important and the most consistently absent, because an honest statement of the boundary of one's own testing is what converts documentation from a marketing artifact into a specification. A general-purpose producer cannot state an envelope for uses it cannot foresee. It can always state the limits of its own knowledge, and that statement is both achievable and, for the integrator, decision-relevant.

For capability-intrinsic risks, the class identified in Section 3.6 and realized by access rather than by deployment design, the duty is not specification at all but evaluation and, where appropriate, restriction on release. The framework does not collapse these into a single obligation, and it is a virtue of allocation by stage rather than by commercial role that it need not.

5.8 The framework applied

The preceding analysis can be reduced to a matrix. Its columns are the capacities Section 3.1 distinguished, with the temporal dimension folded into the control column, and the duty that follows from their intersection. Its rows follow the stages of Section 2, with the agent harness of Section 2.10 entered separately because it introduces an actor the stage model does not otherwise name.

Stage	Knowledge	Effective control	Proportionate duty
Corpus	Producer only	Producer, until collection closes	Characterize population, period, provenance and legal basis; report measured memorization propensity
Architecture	Producer	Producer, until training begins	Disclose model family and interpretability class
Training run	Producer only	Producer; irreversibly closed at completion	Disclose objective, scale, and measured envelope by task and by subgroup
Post-training	Whoever adapts	Whoever adapts; cheap and untraceable	Re-specify the adapted artifact; declare deltas from base, including changes to refusal behavior
Model artifact	Producer	Producer at release, thereafter nobody	Standardized, comparable, mandatory specification sheet; artifact identity by digest, format and precision; license terms
Inference runtime	Integrator or serving provider	Same party; fully reversible	Declare runtime, chat template, sampling defaults, quantization and context policy
Orchestration and application	Integrator only	Integrator; fully reversible	Duty of selection: demonstrate envelope covers population, task and operating point; combine with own base rate; log decisions
Agent harness and platform	Platform knows the tooling; author knows the purpose	Author, within limits the platform sets	Platform: specify the tooling and constrain what can be assembled; author: duty of selection, displaced by constraint where competence is absent
Interface and oversight	Integrator only	Integrator; fully reversible	Demonstrate that oversight is effective rather than nominal
Deployment topology	Integrator chooses; serving provider knows	Both, jointly	Obtain and retain an attested pipeline description; auditability commensurate with risk class

Two properties of the matrix carry the argument. First, the duty in each row is one the named party can actually perform, which is not true of any current instrument read row by row. Second, no row assigns responsibility for *risk* to a party above the application layer, and no row relieves a party above it of responsibility for *statement*. That is the allocation the chapter has been arguing for, and it is what the phrase "risk is knowable only at the application layer" amounts to in operative terms.

5.9 The AI Act as enacted

The proposal is best understood as a completion of a movement the European instrument begins and does not finish.

The AI Act already recognizes a value chain. It distinguishes providers from deployers, importers and distributors; it provides that a party making a substantial modification assumes provider obligations; it requires that information flow to downstream providers who integrate general-purpose models; and it exempts certain freely licensed releases from parts of the regime.[42] These are not the provisions of a statute that imagines a single monolithic system with a single author.

Four gaps remain, and each maps onto a finding above.

Obligations are allocated by *commercial role* rather than by technical stage. Section 3.1 showed that contract is a poor proxy for control: two firms in identical contractual positions may occupy wholly different positions in the control structure depending on which stages each performs. Role-based allocation therefore misfires even where the chain is correctly mapped.

The information that must flow downstream is specified in *substance but not in*

---

[42] Regulation (EU) 2024/1689, *supra* note 1, arts. 2(12), 25, 53-55 and Annex XII. Article 53(1)(b)(ii) refers to Annex XII, which specifies the minimum transparency information for downstream providers integrating a general-purpose AI model. The Regulation has since been amended by the Digital Omnibus on AI, adopted in June 2026 and in force from July 2026, which defers the Chapter III high-risk obligations to 2 December 2027 for stand-alone Annex III systems and to 2 August 2028 for Annex I systems. The amendments adjust timing and simplify particular requirements; they leave the role-based allocation discussed here intact.

form. It is not standardized, not comparable across producers, and not subject to a defined test method, which means it fails three of the four conditions that make the speed rating function.

The exemptions for freely licensed models *relieve* rather than *convert*. Section 4.2.3 argued that open release eliminates the developer's causal access while leaving its epistemic access intact and exclusive, and that the correct response is therefore to transform the duty into one of specification rather than to withdraw it.

And the instrument does not distinguish *reversible from irreversible* stages. Its post-market monitoring and corrective-action architecture is well designed for the application and runtime layers and largely inapplicable to pre-training, where the only performable duty is accurate statement in advance. Section 3.3 argued that this is a physical property of the process rather than a drafting oversight, and that a regime which does not recognize it will demand the impossible upstream while underspecifying the tractable downstream.

One further observation belongs here, because it shows that the difficulty is not an inherent hardness in the subject matter. When the bank of Section 1 replaces its fitted ensemble with a committee-authored tree, it leaves the AI Act, and it also leaves the AI Act's own right to an explanation of an individual decision, which is keyed to high-risk systems and lapses with them.[43] What it does not leave is the general law of automated decision-making. The European data protection regime confers rights in respect of decisions taken solely by automated means that produce legal effects or similarly significantly affect a person, and the automatic refusal of a credit application is its paradigm case. The Brazilian regime confers an analogous right of review and names the definition of a credit profile expressly.[44] Neither asks whether the logic was

---

[43] Regulation (EU) 2024/1689, *supra* note 1, art. 86. Article 86 applies to an affected person subject to a decision taken by a deployer on the basis of the output of a high-risk AI system listed in Annex III, except systems under point 2, where the decision produces legal effects or similarly significant adverse effects on health, safety or fundamental rights. It grants clear and meaningful explanations of the AI system's role and the main elements of the decision, subject to statutory exceptions and only where Union law does not otherwise provide the right.

[44] Regulation (EU) 2016/679 (General Data Protection Regulation), art. 22 and Recital 71 (citing the automatic refusal of an online credit application); Lei No. 13.709/2018 (Lei Geral de Proteção de Dados Pessoais), art. 20.

fitted or authored, because neither takes the technique as its subject. Each takes the decision, the person subject to it, and the effect upon that person.

The contrast is instructive and slightly uncomfortable. Data protection law, drafted before the present generation of models existed, selected a subject that survives every change of technique. Artificial intelligence law, drafted with those models in view, selected a subject that a change of technique dissolves. The lesson is not that the AI Act should be replaced by data protection law, which governs a narrower field and answers different questions; a right to review a decision says nothing about the training corpus of a model that has not yet decided anything, and the specification duty proposed above has no counterpart there. *The lesson is that correct anchoring already exists in the legal order, that it was available, and that it was not followed*.

None of these requires abandoning the instrument's architecture. Each is a refinement of allocation within it, which is, on the argument of this chapter, the only kind of reform that could work, because the object the instrument was drafted around does not exist and the stages that do exist are already, if incompletely, visible in its text.

CONCLUSION

The argument of this chapter has been that the object the world's artificial intelligence instruments were drafted around does not exist. There is no AI system in the sense the definitions require: no single artifact, produced by one identifiable party, whose properties can be assessed and whose risks can be attributed. There is a sequence of stages, separated in time, performed by different parties with different knowledge and different powers, and assembled into deployments whose risk profiles bear no fixed relation to any component.

What follows from this is not a preference for a more granular regulatory architecture. It is that a layered attribution is a condition of any regulation being applicable at all, and the reason is closer to a conceptual constraint than to a policy judgment. Distinct duties can only attach to distinct objects. Where a single obligation is directed at a plurality of artifacts whose relevant properties are held by different parties at different times, the obligation has no determinate addressee: it is discharged by whoever finds it cheapest to document and by no one who could have affected the

outcome. The failure is therefore not that current instruments chose the wrong level of detail. It is that they identified a subject that cannot bear a duty, and no amount of subsequent refinement in the obligations can repair a defect in the thing they are imposed upon.

The calibration this requires is narrower than it may appear. A statute need not enumerate nine stages, nor legislate in the vocabulary of tokenizers and decoding parameters; a legal system asked to absorb that degree of technical specificity would apply it badly, and would be right to resist. But it cannot operate with one object either. The workable level lies between, and Sections 3 and 5 identified it: the law needs to distinguish the actor or set of actors that determines what an artifact can do from the actor or set of actors that determines what it will be used for, to impose on the first a duty of accurate and comparable specification, and on the second a duty of fit. Two duties, two classes of addressee, and a criterion (knowledge joined to control at a moment when action remains possible) for deciding which applies to whom.

One immediate consequence is that a category the European instrument found itself compelled to invent becomes unnecessary. General-purpose artificial intelligence is a repair, and its existence is a symptom. The original architecture keyed obligations to a system's intended purpose, and it therefore had no place for an artifact possessing none. On the account developed here, no model has an intended purpose. Purpose is not a property that models have and general-purpose models lack; it is determined at the application layer in every case, for a credit-scoring regression exactly as for a language model. The distinction between purpose-specific and general-purpose artifacts consequently dissolves, and with it the need for a separate chapter to govern the residue.

The history is instructive on this point, because it shows what the alternative costs. The Commission's original proposal contained no provisions on general-purpose models. The public release of a conversational assistant in late 2022 obliged the Council and Parliament to introduce them, and the resulting chapter was settled under time pressure in the closing negotiations.[45] A regulatory architecture in which

[45] The Commission's proposal of 21 April 2021, COM(2021) 206 final, contained no chapter on general-purpose models; provisions on foundation models and general-purpose artificial intelligence were introduced during the subsequent legislative process and settled in the final negotia-

the arrival of one product requires a new chapter is an architecture built on the wrong object. Nothing about that product was structurally novel: it had a corpus, an architecture, a training run, post-training, an artifact, a runtime, an application, and a deployment topology, exactly as its predecessors did. What was novel was its breadth of capability, and breadth of capability is a parameter of a specification, not a reason for a separate legal category.

Under a layered treatment the three worked systems of this chapter, and the wider families they stand for, are handled by one apparatus. A financial scoring model, a facial embedding model, and a large language model are all components with performance envelopes. They differ in how broad those envelopes are and in how much of the envelope can be characterized in advance, which is a matter of degree and is exactly what Section 5.7 proposed to scale the specification duty against. None of them requires its own regime, and none carries a complete deployment-risk classification at that level; what can be classified upstream are capability envelopes and capability-intrinsic risks. Context-dependent risk arrives with purpose, population, authority, and consequence, and all four are settled above.

That is also where liability belongs, and the point is becoming urgent rather than theoretical. Section 2.10 described what an agent is: a model, unchanged, placed in a loop by software that grants it credentials, tools, memory, and the authority to act without a person in the path. In such a deployment the harness is the actuator. If liability tracks control, it must fall on the party that built or configured the harness and on the platform that determined what could be built, and it must fall there whether the model was licensed, hosted, or served through an interface, because none of those arrangements altered who decided that an action would execute unattended. A regime that continues to locate responsibility in the model will find, as agents proliferate, that it has assigned liability to the only party in the chain that did not act.

---

tions of 2023. The Council adopted its general approach on 6 Dec. 2022. The European Parliament's negotiating position of 14 June 2023 introduced express definitions and obligations for foundation models and general-purpose AI systems. Parliament and Council reached political agreement on 8 Dec. 2023, and the final Regulation recast the category around general-purpose AI models, including a separate regime for models with systemic risk.

The durability of the framework rests on a modest observation. The stages of Section 2 are not features of any particular technique. Whatever succeeds the transformer will still be trained on a corpus that samples some population, under an objective that defines what counts as better, producing an artifact that must be executed by software, integrated into an application that fixes a purpose, and hosted somewhere that determines what can be observed. A regime organized around those stages does not require amendment when the technique changes. A regime organized around a named technology requires amendment every time one arrives, and will always arrive late.

Much remains to be worked out, and this chapter has not attempted it. The test methods that would make model specifications comparable exist for biometrics and barely exist elsewhere. The competence threshold at which a duty of selection gives way to a duty of constraint, identified in Section 5.6, needs content. And the relationship between the specification duty proposed here and the existing consumer and product liability regimes of individual jurisdictions, including the Brazilian one, requires the doctrinal treatment that a chapter on foundations can only gesture toward. But the prior question is the one addressed here, and it is not a question of degree. Regulation cannot allocate responsibility across a process it has declined to look inside.

REFERENCES

AI@META. *Llama 3 Model Card*. 2024.

BRAZIL. Lei No. 8.078, de 11 de setembro de 1990 (Código de Defesa do Consumidor).

BRAZIL. Lei No. 13.709, de 14 de agosto de 2018 (Lei Geral de Proteção de Dados Pessoais).

BRAZIL. Projeto de Lei No. 2338, de 2023. Substitutive approved by the Federal Senate on 10 Dec. 2024 and transmitted to the Chamber of Deputies on 17 Mar. 2025; pending consideration in the Chamber as of 30 July 2026.

CARLINI, Nicholas et al. *Extracting Training Data from Large Language Models*. In: 30th USENIX Security Symposium, p. 2633-2650, 2021.

CARLINI, Nicholas et al. *Quantifying Memorization Across Neural Language Models*. In: International Conference on Learning Representations, 2023.

CENTRO DE ESTUDOS DE SEGURANÇA E CIDADANIA. *O Panóptico: monitor do reconhecimento facial no Brasil*. Rio de Janeiro: CESeC, 2019-.

COBBE, Jennifer; LEE, Michelle Seng Ah; SINGH, Jatinder. *Reviewable Automated Decision-Making: A Framework for Accountable Algorithmic Systems*. In: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, p. 598-609, 2021.

COBBE, Jennifer; VEALE, Michael; SINGH, Jatinder. *Understanding Accountability in Algorithmic Supply Chains*. In: Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency, p. 1186-1197, 2023.

DEEPSEEK. *DeepSeek-R1*: model weights and repository, released under the MIT License. 2025.

DEFENSORIA PÚBLICA DA UNIÃO; CENTRO DE ESTUDOS DE SEGURANÇA E CIDADANIA. *Mapeando a Vigilância Biométrica: levantamento nacional sobre o uso do reconhecimento facial na segurança pública*. Rio de Janeiro: CESeC/DPU, 2025.

DENG, Jiankang et al. *ArcFace: Additive Angular Margin Loss for Deep Face Recognition*. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, p. 4690-4699, 2019.

EUROPEAN COMMISSION. *Guidelines on the definition of an artificial intelligence system established by Regulation (EU) 2024/1689*. C(2025) 924 final, 6 Feb. 2025.

EUROPEAN COMMISSION. *Proposal for a Regulation laying down harmonised rules on artificial intelligence*. COM(2021) 206 final, 21 Apr. 2021.

EUROPEAN UNION. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data (General Data Protection Regulation). Official Journal of the European Union, L 119, 2016.

EUROPEAN UNION. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence. Official Journal of the European Union, L 1689, 2024.

GEBRU, Timnit et al. *Datasheets for Datasets*. Communications of the ACM, v. 64, n. 12, p. 86-92, 2021.

GROTHER, Patrick; NGAN, Mei; HANAOKA, Kayee. *Face Recognition Vendor Test (FRVT) Part 3: Demographic Effects*. NISTIR 8280. Gaithersburg: National Institute of Standards and Technology, 2019.

HU, Edward J. et al. *LoRA: Low-Rank Adaptation of Large Language Models*. In: International Conference on Learning Representations, 2022.

INTERNATIONAL ORGANIZATION FOR STANDARDIZATION; INTERNATIONAL ELECTROTECHNICAL COMMISSION. *ISO/IEC 22989:2022 — Information Technology, Artificial Intelligence, Artificial Intelligence Concepts and Terminology*. Geneva: ISO, 2022.

LECUN, Yann et al. *Gradient-Based Learning Applied to Document Recognition*. Proceedings of the IEEE, v. 86, n. 11, p. 2278-2324, 1998.

LEVESON, Nancy G. *Engineering a Safer World: Systems Thinking Applied to Safety*. Cambridge: MIT

Press, 2011.

META. *Llama 3 Community License Agreement*. 2024.

MITCHELL, Margaret et al. *Model Cards for Model Reporting*. In: Proceedings of the Conference on Fairness, Accountability, and Transparency, p. 220-229, 2019.

NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY. *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*. NIST AI 100-1. Gaithersburg: NIST, 2023.

NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY. *Face Technology Evaluations: FRTE / FATE*. Gaithersburg: NIST, ongoing.

OPENAI. *GPT-4 Technical Report*. arXiv:2303.08774, 2023.

OPEN SOURCE INITIATIVE. *The Open Source AI Definition, version 1.0*. 2024.

ORGANISATION FOR ECONOMIC CO-OPERATION AND DEVELOPMENT. *Recommendation of the Council on Artificial Intelligence*. OECD/LEGAL/0449. Adopted 22 May 2019, amended 3 Nov. 2023.

PARASURAMAN, Raja; RILEY, Victor. *Humans and Automation: Use, Misuse, Disuse, Abuse*. Human Factors, v. 39, n. 2, p. 230-253, 1997.

PERROW, Charles. *Normal Accidents: Living with High-Risk Technologies*. New York: Basic Books, 1984.

ROYAL COMMISSION INTO THE ROBODEBT SCHEME. *Report*. Canberra: Commonwealth of Australia, July 2023.

SELBST, Andrew D. et al. *Fairness and Abstraction in Sociotechnical Systems*. In: Proceedings of the Conference on Fairness, Accountability, and Transparency, p. 59-68, 2019.

SENNRICH, Rico; HADDOW, Barry; BIRCH, Alexandra. *Neural Machine Translation of Rare Words with Subword Units*. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, p. 1715-1725, 2016.

THE AMERICAN LAW INSTITUTE. *Restatement (Third) of Torts: Products Liability*. St. Paul: American Law Institute, 1998.

UNITED NATIONS ECONOMIC COMMISSION FOR EUROPE. *Regulation No. 30 — Uniform Provisions Concerning the Approval of Pneumatic Tyres for Motor Vehicles and Their Trailers*.

VASWANI, Ashish et al. *Attention Is All You Need*. In: Advances in Neural Information Processing Systems 30, 2017.

ARTIFICIAL INTELLIGENCE GOVERNANCE ARCHITECTURES: A CRITICAL ANALYSIS OF THE EUROPEAN AI ACT, THE SOFTWARE-AGNOSTIC APPROACH, AND A HYBRID REGULATORY MODEL BASED ON RISKS AND EVIDENCE

Lucas Martins de Freitas Junior

Librarian, lawyer, and researcher; Master's student in Information Science at UFMG.

DOI: https://doi.org/10.59224/dti8.ch2

Abstract: The expansion of artificial intelligence (AI) has intensified the debate over regulatory models capable of reconciling innovation, legal certainty, and the protection of fundamental rights. This article critically examines AI-specific regulation, represented by the European AI Act, and the approach based on technological neutrality. Through bibliographic and documentary research, it concludes that AI governance requires a hybrid regulatory architecture integrating constitutional principles, risk management, transparency, human oversight, and accountability.

Keywords: digital governance; artificial intelligence regulation; technological neutrality.

Table of contents: 1 Introduction; 2 The crisis in defining artificial intelligence: between technology, law, and regulation; 3 ARTIFICIAL INTELLIGENCE AS A DISTINCT TECHNOLOGY OR SIMPLY ANOTHER FORM OF SOFTWARE?; 4 The european union's horizontal paradigm: the AI act and risk-based governance; 5 The effectiveness problem: predictive AI, technological hype, and the AI snake oil phenomenon; 6 A proposal for hybrid regulatory governance of artificial intelligence;

6.1 First layer: general principles applicable to all digital systems; 6.2 Second layer: regulation based on the context of use; 6.3 Third layer: specific requirements for artificial intelligence systems; 6.4 Fourth layer: evidence-based governance; 6.5 A regulatory architecture guided by proportionality; 7 The construction of artificial intelligence governance in brazil: between international influence and the specific features of the national legal system; 7.1 The Federal Constitution as the foundation of Artificial Intelligence governance; 7.2 The General Data Protection Law and the consolidation of digital governance; 7.3 The Brazilian Artificial Intelligence Strategy and the construction of a national regulatory framework; 7.4 The role of public institutions in constructing Artificial Intelligence governance; 7.5 Challenges to consolidating Artificial Intelligence governance in Brazil; 7.6 Brazilian Artificial Intelligence governance in comparative perspective; 8 Conclusions; References.

1 INTRODUCTION

Artificial intelligence has become one of the main technological infrastructures of contemporary society, influencing economic, scientific, administrative, and legal activities. Systems capable of learning patterns, generating content, supporting decisions, and automating processes have been incorporated into applications in healthcare, education, the financial system, public security, and the Judiciary, profoundly transforming the relationship between technology, society, and law. This scenario has shifted the regulatory debate from data protection to the governance of systems capable of producing decisions with a high degree of autonomy.

The rapid expansion of these technologies has revived the discussion about the need for a specific legal framework for artificial intelligence. Traditionally, the principle of technological neutrality prevailed, according to which the law should regulate the effects produced by technologies rather than the technologies themselves. However, the European Union's adoption of the Artificial Intelligence Act (AI Act) inaugurated a new paradigm by establishing a horizontal regulatory model based on the classification of risks arising from the use of artificial intelligence.

This regulatory choice, however, is not consensual. Part of the literature argues that artificial intelligence does not represent a sufficiently significant technological

rupture to justify its own legal regime and advocates a technology-agnostic approach focused on the risks produced by software systems, regardless of the techniques employed. In addition, a conceptual difficulty persists concerning the legal definition of artificial intelligence itself, since definitions that are too narrow tend to become rapidly obsolete, whereas excessively broad definitions ultimately encompass technologies traditionally governed by other legal instruments.

The advance of generative artificial intelligence has also broadened the debate concerning the effectiveness of these technologies. Recent studies, such as those by Narayanan and Kapoor, show that several applications marketed as artificial intelligence solutions perform below advertised levels, especially in areas related to predicting human behavior, credit, recruitment, and public security. These limitations reinforce the need for regulation to consider not only the technology's potential risks but also its empirically demonstrated capabilities.

Against this background, this article seeks to answer the following research question: does artificial intelligence require specific regulation, as proposed by the European AI Act, or would a technology-agnostic regulatory approach focused on the uses and risks produced by software systems be more appropriate?

The hypothesis is that neither perspective, considered in isolation, is sufficient to address the complexity of contemporary digital governance. It is argued that constructing an effective regulatory model requires the integration of general legal principles, risk management, transparency, human oversight, and specific mechanisms applicable to higher-impact artificial intelligence applications.

This qualitative, exploratory, and bibliographic study is grounded in a critical analysis of national and international literature and normative documents. It engages with authors such as Bart Custers, Joshua Ellul, Giusella Finocchiaro, and Arvind Narayanan, while also examining instruments including the AI Act, Brazil's General Data Protection Law, Bill No. 2,338/2023, and other digital governance frameworks. Finally, it proposes a hybrid regulatory architecture capable of reconciling technological innovation, legal certainty, and the protection of fundamental rights.

2 THE CRISIS IN DEFINING ARTIFICIAL INTELLIGENCE: BETWEEN TECHNOLOGY, LAW, AND REGULATION

The development of any legal regime presupposes the existence of a relatively stable regulatory object. The law must precisely identify what will be subject to legal discipline, delimiting its scope, addressees, and the consequences arising from the application of its rules. This premise, however, encounters a significant obstacle when applied to artificial intelligence. Unlike traditional legal institutions, AI is neither a univocal concept nor a technology with universally accepted technical boundaries, a circumstance that makes its definition one of the greatest challenges of contemporary digital governance.

Although the expression "artificial intelligence" has been used since the Dartmouth Conference in 1956, its development has been extremely heterogeneous. Expert systems, statistical models, machine-learning techniques, deep neural networks, generative systems, symbolic algorithms, natural language processing, computer vision, and hybrid architectures coexist under the same label. This diversity makes it difficult to construct a legal concept precise enough to support regulatory activity without rapidly becoming obsolete in light of the speed of technological innovation. Finocchiaro notes that defining artificial intelligence remains an extremely complex undertaking and that, despite the many proposals formulated over recent decades, none has achieved universal acceptance. As the author states, "defining artificial intelligence is a difficult endeavour, and in fact, many definitions have been proposed" (Finocchiaro, 2023, p. 1961).[1]

This difficulty arises largely from the fact that the concept of intelligence itself remains contested in fields such as philosophy, cognitive psychology, neuroscience, and computer science. Consequently, any attempt to define artificial intelligence in legal terms imports theoretical assumptions that are not always consensual. Rather than seeking the essence of intelligence, Alan Turing proposed shifting the analysis toward the results produced by a system. From this perspective, if a machine can

[1] FINOCCHIARO, Giusella. *The Regulation of Artificial Intelligence*. Cheltenham: Edward Elgar Publishing, 2024, p. 1961.

perform tasks that would be considered intelligent when carried out by human beings, its behavior may be treated as functionally intelligent. For Finocchiaro, this remains the most consistent formulation presented to date precisely because it avoids debating an abstract concept of intelligence and focuses instead on the systems' operational capabilities.

The adoption of this functional criterion, however, does not eliminate legal problems. On the contrary, it significantly expands the range of technologies that could potentially be classified as artificial intelligence. Traditional software, automated decision-making systems, statistical algorithms, and conventional automation mechanisms may produce results similar to those generated by modern machine-learning techniques. If all such systems are classified as AI solely because of their outputs, regulation loses the ability to distinguish technologies displaying very different levels of autonomy, adaptability, and risk.

This phenomenon reveals a recurring problem in the literature: the tendency to anthropomorphize artificial intelligence. The very use of the word "intelligence" encourages the perception that these systems possess cognitive capacities similar to those of human beings, favoring mistaken interpretations concerning their autonomy, intentionality, or moral responsibility. Finocchiaro warns that the indiscriminate use of this metaphor may have significant legal consequences because it leads interpreters to attribute subjective characteristics to systems that, in reality, perform mathematical operations on datasets. Although useful for communication purposes, the metaphor of intelligence cannot replace a technical analysis of existing computational limitations.

Bart Custers (2022) [2] approaches this issue from another perspective. For the author, the principal novelty of artificial intelligence lies not in the mere automation of tasks but in the combination of two specific features: learning capacity and decision-making autonomy. Unlike traditional algorithms, which execute predetermined instructions, machine-learning systems may modify their own models based on data received during operation. This feature partially disrupts the predictability typical of conventional software and explains why many authors support the need

[2] CUSTERS, Bart; FOSCH-VILLARONGA, Eduard (org.). *Law and Artificial Intelligence*. Cham: Springer, 2022.

for regulatory instruments specifically designed for AI applications.

According to Custers, this autonomy does not mean absolute independence or artificial consciousness, but rather the ability to select strategies to achieve objectives previously defined by human beings. In applications such as autonomous vehicles, medical systems, or legal decision-support models, the path adopted by the algorithm is not entirely programmed by the developer. It emerges from a learning process carried out on large volumes of data, which makes both future behavior and the traditional attribution of legal responsibility more difficult to predict.

This distinction between conventional software and artificial intelligence is precisely the main argument advanced by proponents of specific regulation. If certain applications exhibit unprecedented levels of opacity, continuous adaptation, and operational autonomy, subjecting them only to rules traditionally applicable to software would be insufficient. Additional requirements concerning data governance, technical documentation, human oversight, risk assessment, and explainability mechanisms arise precisely in response to these new characteristics.

This argument, however, faces strong opposition in part of the literature. Joshua Ellul maintains that attempting to create a distinct legal category for artificial intelligence generates more problems than solutions. He observes that no sufficiently precise definition can permanently distinguish AI systems from other computational systems. Any concept adopted tends to produce two equally problematic consequences: including technologies that should not be regulated or excluding potentially dangerous technologies merely because they use computational techniques different from those contemplated by the legislation.

For Ellul (2022) [3], the regulatory focus should shift from the technology to the activity being performed. The legally relevant issue would not be whether a given algorithm uses deep learning, symbolic logic, or traditional programming rules, but rather what concrete impacts the system produces on fundamental rights, safety, health, competition, or consumer protection. From this perspective, an automated system capable of discriminating against candidates in a selection process should

[3] ELLUL, Joshua. *Should We Regulate Artificial Intelligence or Some Uses of Software?* In: CUSTERS, Bart; FOSCH-VILLARONGA, Eduard (org.). *Law and Artificial Intelligence*. Cham: Springer, 2022.

receive equivalent legal treatment regardless of whether it employs artificial intelligence or simple statistical rules. Technological neutrality would preserve regulatory stability amid the rapid evolution of computational methods and reduce the risk of legislative obsolescence.

It can therefore be observed that the contemporary debate goes beyond the mere formulation of a technical definition. The true conflict lies in the regulatory architecture itself. Whereas authors such as Custers understand that certain characteristics of artificial intelligence justify specific normative mechanisms, Ellul argues that these characteristics are not exclusive to AI and may also emerge in complex software ecosystems. Finocchiaro, in turn, moves the discussion to a prior level by demonstrating that every regulatory attempt inevitably depends on defining the object being regulated, a definition that remains open and subject to continuous revision.

This controversy demonstrates that defining artificial intelligence is not merely a terminological problem. It is a political, legal, and regulatory decision that conditions the entire digital governance model. Depending on the conception adopted, the scope of the rules is expanded or restricted, regulatory costs change, the duties imposed on developers are modified, and mechanisms for protecting fundamental rights are redefined. For this reason, understanding the conceptual crisis of artificial intelligence is the first step toward critically evaluating the regulatory models currently in dispute, particularly the horizontal paradigm established by the European AI Act and the agnostic approach based on regulating software uses.

3 ARTIFICIAL INTELLIGENCE AS A DISTINCT TECHNOLOGY OR SIMPLY ANOTHER FORM OF SOFTWARE?

The discussion concerning the regulation of artificial intelligence can only be understood by addressing a prior question: does AI effectively constitute a distinct technological category, or is it merely another development in software engineering? The answer determines not only how legal rules are drafted but also the institutional architecture responsible for digital governance. In other words, before deciding how to regulate artificial intelligence, it is necessary to determine whether sufficient technical grounds exist to treat it as an autonomous legal object.

For decades, technology law was developed under the principle of technological

neutrality. This principle rests on the premise that legislation should not be designed to regulate specific technologies, but rather the behaviors, activities, or effects produced by those technologies. Rules concerning civil liability, data protection, consumer protection, or competition therefore remain valid regardless of changes in computational development. This feature contributes to regulatory stability and prevents rapid technological evolution from rendering the legal system obsolete within a short period.

The emergence of machine-learning-based artificial intelligence systems, however, prompted questions about the adequacy of this traditional model. According to part of the literature, AI introduced characteristics qualitatively different from those found in conventional software, especially because of its capacity for learning, continuous adaptation, and operational autonomy. From this perspective, artificial intelligence would cease to be merely a tool programmed to execute predefined instructions and would instead constitute a computational architecture capable of building its own decision models from experience accumulated during use.

Bart Custers (2022) [4] argues that this transformation represents a significant rupture in the evolution of digital technologies. According to the author, unlike traditional algorithms, whose logic remains entirely determined by the programmer, machine-learning systems develop statistical patterns capable of modifying their behavior without direct human intervention. In this context, autonomy does not mean absolute independence, but the capacity to select different strategies to achieve previously established objectives. As Custers states, "AI systems can make decisions themselves, decisions that are not preprogrammed" (Custers, 2022, p. 8) [5]. This emergent behavior produces significant legal consequences, particularly because it reduces the predictability of decisions and complicates the traditional attribution of responsibility to system developers, operators, and users.

From this perspective, artificial intelligence displays characteristics that go beyond those normally associated with conventional software. Whereas deterministic programs always produce the same result when presented with the same input, machine-learning models may continuously alter their internal parameters as they

[4] CUSTERS, *op. cit.*

[5] *Ibid.*, p. 8.

receive new data during operation. The practical consequence is that a system's future behavior cannot always be fully anticipated even by its own developer. This makes governance mechanisms concerning training-data quality, technical documentation, human oversight, and the explainability of automated decisions particularly relevant.

This interpretation directly influenced the construction of the European regulatory model. By recognizing that certain artificial intelligence applications display unprecedented levels of opacity and autonomy, the AI Act establishes specific obligations that do not apply to traditional software, such as prior conformity assessments, permanent risk-management systems, post-market monitoring, and detailed data-governance requirements. The regulation's very structure rests on the premise that certain technical characteristics justify differentiated legal treatment.

This conclusion, however, is intensely contested. Joshua Ellul (2022) [6] questions precisely the assumption that a sufficiently clear boundary exists between artificial intelligence and conventional software. According to the author, any attempt to construct a specific legal category for AI faces an unavoidable problem: constant technological evolution continuously changes what is understood as artificial intelligence. Techniques regarded as innovative today become ordinary functionalities only a few years later, while new computational architectures continually emerge beyond the classifications anticipated by lawmakers.

Ellul argues that this conceptual instability produces a phenomenon known as a regulatory gap. Whenever legislation seeks to define precisely what constitutes artificial intelligence, it creates room for functionally equivalent systems to remain outside the regulatory scope merely because they use computational methods different from those contemplated in the legal definition. Conversely, excessively broad definitions subject traditional programs to the same regulatory burdens imposed on much more complex systems, increasing compliance costs without proportionate benefits for society.

Ellul's criticism becomes particularly relevant when viewed from the perspective of software engineering. Contemporary digital platforms consist of hundreds or thousands of distributed components, often developed by different teams and simul-

[6] ELLUL, *op cit.*

taneously employing statistical techniques, classical algorithms, external services, and machine-learning models. In such environments, establishing an objective separation between what constitutes artificial intelligence and what merely represents complex software becomes practically impossible. As a result, the application of legal rules depends more on the technical classification adopted by the developer than on the risks actually produced by the system.

This reasoning leads to the main contribution of the agnostic approach. For Ellul, the law should abandon its concern with identifying whether a given system uses artificial intelligence and instead focus on the nature of the activity performed. As the author states, "regulation can be technology agnostic and focus on behaviours and outcomes rather than technology-specifics" (Ellul, 2022, p. 4) [7]. The object of regulation would become the context in which the technology is used and its concrete effects on individuals and organizations. Thus, an automated system used for granting credit, recruiting workers, or making administrative decisions should be subject to the same standards of transparency, auditing, and accountability regardless of whether it employs deep neural networks, traditional statistical models, or preprogrammed rules.

This perspective is closely aligned with the traditional principle of technological neutrality, widely used in several branches of digital law. Instead of regulating specific technologies, it seeks to protect enduring legal values such as equality, security, privacy, due process, and non-discrimination. The advantage of this model lies precisely in its ability to remain applicable despite rapid technological evolution, reducing the need for successive legislative reforms.

The adoption of a fully agnostic model nevertheless presents important limitations. Certain technical characteristics of artificial intelligence—especially algorithmic opacity, dependence on large volumes of data, continuous learning, and the difficulty of explaining decisions—create regulatory challenges that traditional rules designed for conventional software are unlikely to address adequately. Requiring explainability mechanisms for a deterministic algorithm has a completely different meaning from imposing the same requirement on deep neural networks composed of billions of parameters, whose decisions frequently escape even their developers'

[7] *Ibid.* p.*4*

understanding.

At this point, it becomes clear that the two perspectives are not necessarily incompatible. Whereas Custers emphasizes the technical specificities of artificial intelligence to justify tailored regulatory mechanisms, Ellul seeks to preserve normative stability by maintaining technological neutrality. Both, however, share a common concern: preventing technological evolution from creating situations of absent responsibility or weakening the protection of fundamental rights. Their disagreement concerns less the objectives pursued than the regulatory architecture considered most appropriate for achieving them.

From this perspective, the real challenge is not choosing between regulating artificial intelligence and regulating software, but identifying which elements justify differentiated legal treatment. Systems used in contexts of high social impact may require specific governance requirements regardless of the computational technique employed, while certain characteristics intrinsic to machine-learning models justify additional obligations concerning transparency, human oversight, and the quality of the data used for training.

This finding suggests that the apparent opposition between specific regulation and technological neutrality may represent a false dichotomy. The growing complexity of digital ecosystems indicates that contemporary regulatory models will need to combine general principles applicable to all software with specific mechanisms for artificial intelligence applications capable of producing differentiated risks. It is precisely this attempt at reconciliation that guides the European risk-based model, whose structure will be examined in the following section.

4 THE EUROPEAN UNION'S HORIZONTAL PARADIGM: THE AI ACT AND RISK-BASED GOVERNANCE

The debate over artificial intelligence regulation reached its most significant milestone with the European Union's adoption of the Artificial Intelligence Act (AI Act). For the first time, a major jurisdiction adopted a horizontal regulatory framework specifically directed at artificial intelligence, establishing legal obligations that vary according to the level of risk presented by different technological applications. More than sector-specific legislation, the AI Act represents a paradigmatic change in

how the law understands the relationship among technological innovation, the protection of fundamental rights, and digital governance.

The European choice did not arise in isolation. Since the General Data Protection Regulation (GDPR) entered into force, the European Union has consolidated a regulatory strategy based on anticipating the social impacts of digital technologies. Rather than reacting only after concrete harm occurs, European lawmakers have prioritized preventive mechanisms by imposing duties of risk assessment, transparency, technical documentation, and accountability even before certain products are placed on the market. The AI Act continues this regulatory strategy, now directed specifically at artificial intelligence systems.

As Bart Custers (2022) [8] observes, the rapid expansion of artificial intelligence has produced simultaneous transformations in virtually every branch of law. Areas such as civil liability, consumer law, intellectual property, labor law, judicial procedure, competition, data protection, and criminal law have begun to face challenges arising from the growing autonomy of computational systems. For the author, the speed of this transformation justifies the construction of regulatory instruments capable of addressing specific AI characteristics, particularly autonomy, continuous learning, and the ability to produce partially unpredictable decisions.

In this context, the AI Act adopts the risk-based approach as its organizing principle. Unlike models that seek to distinguish permitted and prohibited technologies using purely technical criteria, the European regulation focuses on the consequences that certain systems may produce for fundamental rights, safety, health, and democracy. Accordingly, it is not the mere existence of artificial intelligence that determines the intensity of regulation, but the potential harm associated with its use.

This architecture resembles traditional areas of economic and health regulation. For decades, medicines, aircraft, medical devices, and chemical products have been subjected to procedures proportional to the degree of risk they pose to society. The AI Act transposes this logic to artificial intelligence, seeking to balance two objectives frequently presented as conflicting: fostering technological innovation and preserving fundamental rights.

---

[8] CUSTERS, *op. cit.*

To achieve this balance, the regulation organizes artificial intelligence systems into different risk categories.

The first level comprises unacceptable-risk practices considered incompatible with the European Union's fundamental values. This category includes applications capable of manipulating behavior subliminally, exploiting vulnerabilities resulting from age or disability, conducting certain forms of state-sponsored social scoring, and other uses deemed incompatible with the protection of human dignity. In such cases, the legislative choice is not to impose additional requirements but simply to prohibit their use.

The second category consists of high-risk systems, which form the core of the AI Act. It includes applications used in sensitive sectors such as healthcare, education, critical infrastructure, employment relations, recruitment, public administration, law enforcement, migration control, and the administration of justice. Although such systems remain permitted, their use depends on compliance with a rigorous set of obligations concerning continuous risk management, the quality of training data, technical documentation, record-keeping, transparency, human oversight, and monitoring after they are placed on the market.

This regulatory choice reflects an important change in the understanding of technological responsibility. Rather than limiting itself to ex post liability for damage, the AI Act seeks to establish a culture of preventive governance by requiring providers and deployers to demonstrate in advance that they have adopted mechanisms designed to reduce foreseeable risks. This logic resembles the precautionary principle traditionally used in environmental and health law, adapted to the specific features of algorithmic systems.

In addition to high-risk systems, the regulation establishes specific obligations for so-called general-purpose AI models (GPAI), a category introduced in response to the emergence of large foundation models used in applications such as conversational assistants, automated image generation, and multimodal content production. These models possess a feature that distinguishes them from traditional applications: they may be incorporated by thousands of developers in contexts entirely different from those initially envisioned by their creators. Because of this broad capacity for diffusion, the regulation establishes specific duties concerning technical documenta-

tion, transparency, respect for copyright, and the assessment of systemic risks.

The introduction of this category demonstrates the adaptive capacity of the European model. Whereas the earliest versions of the legislative proposal focused predominantly on specific artificial intelligence applications, the rapid evolution of generative AI required the incorporation of new instruments capable of governing technologies that did not even exist when regulatory discussions began. This aspect reveals both the flexibility and the inherent difficulty of drafting rules for technologies in constant transformation.

Precisely because it adopts such a broad regulatory structure, however, the AI Act has become the target of significant criticism. Giusella Finocchiaro (2023) [9] observes that choosing horizontal regulation necessarily presupposes a sufficiently broad definition of artificial intelligence, considerably expanding the universe of systems potentially subject to the new legal requirements. She emphasizes that this model may generate high compliance costs, especially for small and medium-sized enterprises, while favoring large organizations with greater financial and technical capacity to implement complex governance, auditing, and certification processes.

Another aspect frequently highlighted in the literature is the so-called Brussels Effect. The concept describes the European Union's capacity to influence global regulatory standards without resorting to formal extraterritorial mechanisms. Given the economic importance of the European market, multinational companies often choose to apply the requirements established by European legislation throughout their worldwide operations, thereby avoiding the maintenance of different compliance standards for different jurisdictions. The result is the indirect internationalization of European rules, which become a reference for several countries, including those that do not participate in the European legislative process.

From this perspective, the AI Act extends beyond the territorial boundaries of the European Union and begins to influence regulatory debates in different parts of the world. Several national legislative proposals have incorporated concepts such as risk classification, algorithmic impact assessment, human oversight, data governance, and transparency directly inspired by the European model. This movement rein-

[9] FINOCCHIARO, *op.cit.*

forces the European Union's normative leadership in the construction of global artificial intelligence governance.

Nevertheless, the main criticism advanced by proponents of the agnostic approach remains unresolved: does the creation of legislation specifically for artificial intelligence truly represent the best regulatory strategy? If most risks arise from the activity performed rather than from the technology used, a legal regime based exclusively on artificial intelligence may generate regulatory costs greater than the benefits actually achieved.

This question becomes even more relevant in light of the rapid expansion of generative AI and the growing convergence between traditional software techniques and advanced machine-learning models. The boundary between artificial intelligence and conventional software is becoming increasingly blurred, once again bringing the debate on technological neutrality discussed in the previous section to the forefront.

Consequently, although the AI Act represents the most sophisticated regulatory model currently in existence, its adoption does not conclude the debate over the best form of digital governance. On the contrary, it inaugurates a new stage in the discussion, marked by the need to empirically assess whether the costs imposed by specific regulation will be offset by an effective reduction in the risks associated with the use of artificial intelligence.

5 THE EFFECTIVENESS PROBLEM: PREDICTIVE AI, TECHNOLOGICAL HYPE, AND THE AI SNAKE OIL PHENOMENON

The consolidation of artificial intelligence as an object of intense regulatory activity occurred alongside unprecedented growth in public attention devoted to algorithmic technologies. Advances in large language models, generative image systems, and deep-learning applications significantly expanded society's perception of artificial intelligence capabilities. This perception, however, does not always correspond to the performance actually observed in concrete applications. In many cases, excessively optimistic expectations began to guide political decisions, private investments, and regulatory initiatives, creating an environment in which technological promises

frequently precede robust scientific evidence.

In this context, Arvind Narayanan and Sayash Kapoor (2024) [10] offer a fundamental critique of the contemporary debate on artificial intelligence. According to the authors, much of the public discussion treats AI as a homogeneous technology, ignoring substantial differences among its various forms and applications. This simplification leads both to overestimating and underestimating the risks associated with computational systems, making it more difficult to design public policies proportionate to these technologies' actual capabilities.

The authors initially distinguish two broad groups of applications. The first is predictive artificial intelligence, whose purpose is to estimate future events or infer individual characteristics from large datasets. This category includes algorithms used to predict criminal recidivism, default risk, worker productivity, school dropout, academic performance, disease probability, or consumer behavior. The second group comprises generative artificial intelligence, designed to produce texts, images, computer code, audio, or other synthetic content. Although both employ machine-learning techniques, their objectives, limitations, and legal risks differ significantly.

This distinction has important regulatory consequences. Whereas generative systems raise concerns related to disinformation, intellectual property, content authenticity, and information manipulation, predictive applications frequently interfere directly with administrative, financial, or judicial decision-making that affects individual rights. Consequently, treating all these technologies as a single normative category may lead to the adoption of regulatory mechanisms ill-suited to profoundly different technological realities.

It is precisely in this context that Narayanan and Kapoor introduce the concept of AI Snake Oil. As the authors define it, "AI snake oil is AI that does not and cannot work as advertised" (Narayanan; Kapoor, 2024, p. 2) [11]. The expression historically refers to the so-called snake oil salesmen, nineteenth-century American merchants who advertised miracle medicines without scientific evidence. According to the authors, a significant portion of the contemporary artificial intelligence market

---

[10] NARAYANAN, Arvind; KAPOOR, Sayash. *AI Snake Oil: What Artificial Intelligence Can Do, What It Can't, and How to Tell the Difference*. Princeton: Princeton University Press, 2024.

[11] Ibid., p. 2.

reproduces a similar dynamic by marketing technological solutions whose capabilities are systematically exaggerated or presented without sufficient empirical validation. Their criticism is directed primarily at applications based on predicting human behavior. Various systems promise to identify future criminals, predict professional performance, estimate personality, detect emotions, or anticipate individual decisions based on behavioral data. However, numerous studies show that such predictions have high error rates, limited generalizability, and strong sensitivity to the quality of the training data. In many cases, the performance achieved does not surpass traditional statistical methods, even though it is marketed as an exclusive result of artificial intelligence.

From a legal perspective, this phenomenon produces a particularly significant consequence: the mere use of artificial intelligence does not guarantee greater reliability in automated decisions. On the contrary, incorporating complex algorithms may create a false appearance of scientific objectivity, leading public officials, judges, companies, and citizens to place a high degree of trust in results whose accuracy remains limited. This phenomenon, often referred to as automation bias, is the human tendency to accept automated recommendations without subjecting them to the same degree of scrutiny applied to human decisions.

This criticism is especially important for the law. Several contemporary artificial intelligence applications are used precisely in contexts where errors have significant consequences for fundamental rights. Credit-scoring systems may restrict access to financial services; recruitment algorithms may reproduce discriminatory patterns; tools used in public security may expand selective policing practices; and applications used in the administration of justice may influence decisions affecting individual liberty. In these contexts, the system's statistical reliability becomes an essential element of the legitimacy of the administrative or judicial decision.

By shifting the focus to the empirical effectiveness of technologies, Narayanan and Kapoor make a contribution that both complements and, to some extent, challenges the debates conducted by Custers, Ellul, and Finocchiaro. Whereas these authors concentrate on the legal definition of artificial intelligence and the most appropriate institutional architecture for its regulation, the concept of AI Snake Oil introduces a prior question: is there sufficient evidence that certain applications actually work as advertised?

This question significantly changes the terms of the regulatory debate. If certain applications display poor or scientifically disputed performance, the legal problem may concern not only how they should be regulated but also the very legitimacy of their use in sensitive sectors. In other words, before establishing requirements for transparency, governance, or human oversight, it may be necessary to ask whether certain systems should be used in activities capable of restricting fundamental rights.

This perspective resembles the logic adopted in other regulatory fields with high social impact. In the pharmaceutical sector, for example, medicines may be marketed only after their efficacy and safety have been demonstrated through scientific evidence. Medical devices, aircraft, and chemical products are also subject to rigorous certification procedures before being used on a large scale. The analogy shows that technological innovation alone is not a sufficient basis for justifying adoption in critical contexts.

From this standpoint, artificial intelligence governance comes to involve not only risk-mitigation mechanisms but also permanent procedures for performance assessment, scientific validation, and monitoring the results produced by systems after implementation. Human oversight is no longer understood solely as the possibility of individual intervention in automated decisions; it also includes independent audits, robustness testing, periodic accuracy assessments, and continuous review of the models employed.

This debate also reveals a limitation of regulatory approaches based exclusively on technological classification. Two systems classified as artificial intelligence may display completely different levels of reliability, transparency, and social impact. Likewise, applications developed through traditional software techniques may outperform advanced machine-learning models. The intensity of regulatory intervention should therefore depend not only on the technology employed but also on the quality of the available evidence concerning its actual operation.

Consequently, Narayanan and Kapoor's contribution broadens the horizon of digital governance by introducing an element frequently neglected in legal debate: the need to distinguish technological innovation from demonstrated effectiveness. Regulating artificial intelligence systems requires not only understanding their technical architecture or potential risks but also critically assessing the consistency of the

claims made by developers and suppliers. Without this concern, the law risks legitimizing technologies whose effectiveness remains insufficiently demonstrated, thereby compromising both the protection of fundamental rights and the credibility of public policies aimed at digital transformation.

6 A PROPOSAL FOR HYBRID REGULATORY GOVERNANCE OF ARTIFICIAL INTELLIGENCE

The analysis developed in the preceding sections shows that contemporary regulatory debate tends to be presented as a binary choice between two alternatives: on the one hand, creating a legal regime specifically for artificial intelligence, represented by the European AI Act; on the other, maintaining the principle of technological neutrality, defended by authors who propose regulating software uses regardless of the computational techniques employed. Although both perspectives rest on consistent foundations, neither appears sufficient, in isolation, to address the challenges imposed by rapid digital transformation.

Specific regulation of artificial intelligence has the merit of recognizing that certain technical characteristics, such as machine learning, operational autonomy, algorithmic opacity, and continuous adaptation, produce risks that traditional law did not address. Requirements concerning data governance, technical documentation, impact assessment, human oversight, and continuous monitoring represent significant advances in protecting fundamental rights and holding accountable the actors involved in developing and using these systems.

However, as demonstrated by Ellul and Finocchiaro, this model faces limitations arising from the difficulty of objectively and permanently defining what constitutes artificial intelligence. Technological evolution makes every definition potentially temporary, creating risks of regulatory obsolescence and inconsistencies in the application of legislation. Moreover, different systems capable of producing similar impacts may receive different legal treatment solely because of the computational technique used, contrary to the principle of regulatory equality.

By contrast, the software-agnostic approach preserves technological neutrality and reduces dependence on technical concepts subject to rapid change. By focusing regulation on the effects produced by digital applications, this model promotes

greater normative stability and enhances its ability to adapt to future generations of technology. Nevertheless, this perspective also has limitations. Completely disregarding the technical specificities of artificial intelligence may make it difficult to create mechanisms to address issues such as explainability, model training, data quality, model drift, and risks arising from continuous learning.

These findings suggest that the regulatory problem should not be solved by exclusively choosing one of the two paradigms, but by constructing an integrated model capable of combining their principal strengths and reducing their limitations. In this context, a hybrid regulatory-governance architecture structured in four complementary layers is proposed.

6.1 First layer: general principles applicable to all digital systems

The first layer consists of legal principles applicable regardless of the technology employed. It includes rules concerning personal data protection, non-discrimination, administrative transparency, civil liability, information security, consumer protection, due process, and respect for fundamental rights.

The objective of this layer is to preserve the technological neutrality of the legal system. Whenever a computational system produces significant effects on individuals or organizations, it must observe these principles, regardless of whether it uses artificial intelligence, traditional statistical algorithms, or conventional automation mechanisms.

This structure reduces dependence on technical definitions and ensures greater regulatory stability in the face of technological evolution.

6.2 Second layer: regulation based on the context of use

The second layer shifts the focus from the technology to the context in which it is employed. Inspired by Ellul's approach, this dimension is based on the premise that different applications produce different risks even when they use similar technologies.

An algorithm used to recommend films on an entertainment platform does not

have the same potential impact as a system used to select candidates in public examinations, grant bank credit, support judicial decisions, or provide medical diagnoses. Regulatory duties should therefore vary according to the economic sector, the nature of the activity performed, and the intensity of the rights potentially affected.

This logic resembles the existing regulatory tradition in areas such as healthcare, financial markets, air transport, and telecommunications, where the degree of state intervention derives far more from the risk associated with the activity than from the technology employed.

6.3 Third layer: specific requirements for artificial intelligence systems

Although technological neutrality remains the organizing principle, certain characteristics of artificial intelligence justify additional obligations. This third layer incorporates part of the logic developed by the AI Act, limiting its application to systems that effectively display characteristics such as machine learning, a high degree of decisional opacity, continuous adaptation, or significant operational autonomy.

In such cases, requirements concerning documentation of the development lifecycle, traceability of the data used for training, human-oversight mechanisms, independent audits, permanent risk management, and post-deployment monitoring are justified.

The application of these obligations, however, should not arise exclusively from formally classifying the system as artificial intelligence, but from demonstrating that its technical characteristics effectively increase regulatory risks.

6.4 Fourth layer: evidence-based governance

The principal innovation of the proposal is the incorporation of a fourth layer inspired by the contributions of Narayanan and Kapoor. Unlike currently predominant models, this dimension introduces the principle of evidence-based governance.

From this perspective, the intensity of regulatory intervention should depend not only on the type of technology used or exclusively on the context of application, but

also on the quality of the empirical evidence available concerning the system's performance.

Applications intended for healthcare, public security, the administration of justice, education, or public policy should verifiably demonstrate minimum indicators of accuracy, robustness, stability, absence of systematic bias, and operational effectiveness before large-scale implementation.

This mechanism brings artificial intelligence closer to the regulatory regimes traditionally applied to medicines, medical devices, and products with high social impact, in which demonstrating efficacy is an indispensable requirement for authorization of use.

6.5 A regulatory architecture guided by proportionality

The principal advantage of this structure lies in adopting the principle of regulatory proportionality. Instead of imposing the same obligations on all artificial intelligence applications, the model allows the intensity of state intervention to be calibrated according to three complementary factors:

- the risk inherent in the context of use;
- the technical characteristics actually present in the system;
- the scientific evidence available concerning its performance.

This combination reduces unnecessary regulatory costs for low-impact applications, preserves incentives for technological innovation, and simultaneously strengthens protection mechanisms in sectors where algorithmic errors may have serious consequences for individuals and society as a whole.

From this perspective, digital governance is no longer understood as the mere production of rules specifically for artificial intelligence and instead becomes a continuous process of assessment, monitoring, and institutional adaptation. The rapid evolution of digital technologies requires regulatory models capable of incorporating new scientific knowledge without necessarily depending on constant legislative reform.

Consequently, the proposal presented seeks to overcome the traditional opposition between specific regulation of artificial intelligence and technological neutrality.

Rather than choosing one of these paradigms, it argues that a multilayered regulatory architecture provides better conditions for reconciling innovation, legal certainty, the protection of fundamental rights, and responsible technological development.

7 THE CONSTRUCTION OF ARTIFICIAL INTELLIGENCE GOVERNANCE IN BRAZIL: BETWEEN INTERNATIONAL INFLUENCE AND THE SPECIFIC FEATURES OF THE NATIONAL LEGAL SYSTEM

The regulation of artificial intelligence is one of the principal contemporary challenges facing the law, requiring legal systems to reconcile technological innovation, economic development, and the protection of fundamental rights. Although the international debate has been strongly driven by the European Union's adoption of the Artificial Intelligence Act (AI Act) and by guidelines developed by international organizations such as the Organisation for Economic Co-operation and Development (OECD), the United Nations Educational, Scientific and Cultural Organization (UNESCO), and the National Institute of Standards and Technology (NIST), the Brazilian experience has followed its own path, marked by the gradual adaptation of pre-existing legal instruments to the new demands arising from the use of artificial intelligence systems.

Unlike the European experience, characterized by the development of a regulatory framework specifically for artificial intelligence, the Brazilian legal system has consolidated its governance through the integration of constitutional principles, personal data-protection rules, public policies for digital transformation, and regulatory initiatives developed by different state institutions. This is an incremental process in which artificial intelligence was initially governed by legal institutions designed to regulate broader phenomena of digital society, with legislative proposals specifically directed at the particular features of this technology being incorporated at a later stage.

This evolution demonstrates that artificial intelligence governance in Brazil does not result from the simple importation of foreign models. Although the influence of the AI Act and international governance frameworks is undeniable, the national normative structure reflects specific features of Brazilian constitutionalism, especial-

ly the centrality of fundamental rights, personal data protection, and coordinated action among the different branches of government. In this context, the Brazilian experience provides an important subject of analysis for understanding how different regulatory models may be adapted to each country's institutional particularities, reinforcing the hypothesis developed in this study that artificial intelligence governance tends to be structured in hybrid form, reconciling general rules of the legal system with specific mechanisms for risk management and technological oversight.

7.1 The Federal Constitution as the foundation of Artificial Intelligence governance

The 1988 Constitution of the Federative Republic of Brazil constitutes the principal normative foundation of artificial intelligence governance in the country. Although enacted at a historical moment preceding the consolidation of contemporary digital technologies, its organizing principles remain fully applicable to new forms of automation and to the challenges arising from the use of intelligent systems by both the public and private sectors.

Human dignity, established as a foundation of the Republic, sets a substantive limit on the development and use of artificial intelligence, requiring technological innovation to remain subordinate to the protection of individual autonomy, freedom, and equality. Likewise, fundamental rights related to due process, adversarial proceedings, a full defense, intimacy, privacy, and access to justice retain their full effectiveness in the face of the growing use of automated decisions.

The preservation of these rights is consistent with Mireille Hildebrandt's conception of the Digital Rule of Law, according to which the use of algorithmic systems is legitimate only when compatible with the constitutional guarantees that limit the exercise of power. From this perspective, automation neither replaces the statement of reasons for administrative acts nor eliminates the possibility of judicial review of decisions supported by artificial intelligence. On the contrary, the greater the degree of autonomy granted to technological systems, the greater institutions' commitment must be to transparency, human oversight, and accountability.

Within Public Administration, the principles of legality, impersonality, morality, publicity, and efficiency also condition the implementation of artificial intelligence

solutions. The use of these technologies may contribute to improving public-service delivery, streamlining administrative procedures, and strengthening the state's capacity to process information. These benefits, however, do not eliminate the need to observe the constitutional guarantees that structure administrative action. Artificial intelligence therefore remains an auxiliary instrument of state activity, whose legitimacy derives from conformity with constitutional principles rather than from mere technological efficiency.

7.2 The General Data Protection Law and the consolidation of digital governance

The enactment of Law No. 13,709 of August 14, 2018 (General Personal Data Protection Law—LGPD) represents one of the most important milestones in Brazil's digital transformation and constitutes an important foundation for the future regulation of artificial intelligence. Although its principal object is the regulation of personal data processing, its principles and accountability mechanisms directly influence the development, training, and use of intelligent systems, whose operation depends to a large extent on the collection and processing of large volumes of information.

By establishing principles such as purpose limitation, adequacy, necessity, transparency, security, prevention, non-discrimination, and accountability, the LGPD introduced into the Brazilian legal system a governance logic based on the responsible management of risks arising from data processing. These elements have come to guide not only traditional personal-information processing activities but also the development of artificial intelligence applications capable of producing significant impacts on individuals and communities.

The treatment afforded to automated decisions is especially important. Although the interpretation of Article 20 of the LGPD remains the subject of doctrinal debate, its inclusion demonstrates legislative recognition that decisions produced by automated systems require specific mechanisms to protect data subjects. The right to review automated decisions, combined with duties of transparency and accountability, brings Brazilian legislation closer to international discussions concerning human oversight and algorithmic governance.

In this context, Brazilian scholarship emphasizes that personal data protection

goes beyond safeguarding privacy and constitutes an essential instrument for protecting personality and informational self-determination. Authors such as Danilo Doneda, Bruno Bioni, and Laura Schertel Mendes argue that data governance is indispensable to ensuring that the use of artificial intelligence technologies remains compatible with fundamental rights, especially in scenarios characterized by massive information processing and the growing automation of decisions.

7.3 The Brazilian Artificial Intelligence Strategy and the construction of a national regulatory framework

The debate on artificial intelligence in Brazil gained greater institutional density with the publication of the Brazilian Artificial Intelligence Strategy (EBIA), approved in 2021. Although it is not legislative in nature, the strategy represents an important public-policy instrument by establishing guidelines for the scientific, technological, and economic development of artificial intelligence, reconciling incentives for innovation with concerns related to ethics, transparency, security, and the protection of fundamental rights.

The EBIA demonstrates that artificial intelligence governance is not limited to the production of legal rules. Its content shows that the sustainable development of this technology also depends on human-resource training, the strengthening of scientific research, international cooperation, the modernization of Public Administration, and the creation of regulatory environments capable of encouraging innovation without compromising the protection of citizens.

Alongside the development of these public policies, the National Congress began discussing the creation of a specific regulatory framework for artificial intelligence. In this context, Bill No. 2,338/2023 stands out as the principal Brazilian legislative initiative intended to provide general regulation of the subject. Inspired by several aspects of the European AI Act, the bill adopts a risk-based approach, establishes differentiated categories of artificial intelligence systems, and provides specific obligations for applications considered to have greater impact.

The Brazilian bill, however, is not merely a reproduction of the European model. Its structure seeks to make artificial intelligence regulation compatible with institutions already consolidated in the national legal system, particularly the Federal Con-

stitution, the General Data Protection Law, the Brazilian Civil Rights Framework for the Internet, and administrative legislation. This adaptation shows that the Brazilian experience seeks to incorporate international best practices without disregarding the country's constitutional, institutional, and socioeconomic particularities.

7.4 The role of public institutions in constructing Artificial Intelligence governance

The evolution of artificial intelligence governance in Brazil demonstrates that legislative activity is not the only mechanism for producing norms. Several public institutions have begun to play an important role in formulating guidelines, regulations, and practices aimed at the responsible development of artificial intelligence.

The National Data Protection Authority (ANPD), created by the LGPD, has become a strategic actor in interpreting the impacts of artificial intelligence on personal data protection. Although its jurisdiction is not limited to intelligent systems, its guidance directly influences issues related to algorithmic transparency, automated data processing, risk assessment, and the accountability of data-processing agents.

Within the Judiciary, the work of the National Council of Justice (CNJ) stands out. The Council has been developing specific guidelines to regulate the use of artificial intelligence systems by Brazilian courts. CNJ Resolution No. 615/2025 represents an important advance by establishing parameters concerning governance, human oversight, transparency, documentation, security, and risk management applicable to technological solutions used in judicial activity.

The work of these institutions demonstrates that artificial intelligence governance in Brazil has a polycentric nature, characterized by the distribution of responsibilities among different public bodies, regulatory authorities, and institutions responsible for implementing digital-transformation policies. This characteristic is consistent with the conceptions developed by Luciano Floridi and Virginia Dignum (2023) [12], according to which artificial intelligence governance depends on coordination among multiple institutional actors rather than solely on the enactment of legisla-

[12] FLORIDI, Luciano. *The Ethics of Artificial Intelligence: Principles, Challenges, and Opportunities*. Oxford: Oxford University Press, 2023.

tion.

7.5 Challenges to consolidating Artificial Intelligence governance in Brazil

Despite the advances observed, the consolidation of artificial intelligence governance in Brazil remains surrounded by legal, institutional, and technological challenges. The growing incorporation of intelligent systems into public and private activities requires the country to develop mechanisms capable of reconciling legal certainty, the protection of fundamental rights, and incentives for innovation.

One of the principal challenges is harmonizing the different legal regimes applicable to digital transformation. The Federal Constitution, the General Data Protection Law, the Brazilian Civil Rights Framework for the Internet, administrative rules, and future legislation specifically governing artificial intelligence will need to operate in a coordinated manner, avoiding interpretive conflicts and regulatory overlaps that could compromise the effectiveness of governance.

Another relevant aspect concerns strengthening the state's institutional capacity. Implementing algorithmic audits, impact assessments, system certification, and permanent oversight mechanisms requires qualified human resources, adequate technological infrastructure, and cooperation among legal, computational, and administrative specialists. Without such institutional strengthening, the mere enactment of legal rules is likely to produce limited results.

Attention must also be given to structural challenges related to digital inequality, concentration in the technology market, dependence on foreign platforms, and the strengthening of national research and innovation capacity. The construction of an artificial intelligence public policy must go beyond a strictly regulatory debate by incorporating strategies directed toward scientific development, technological sovereignty, and the expansion of Brazil's digital infrastructure.

7.6 Brazilian Artificial Intelligence governance in comparative perspective

Analysis of the Brazilian experience shows that its regulatory evolution differs

significantly from the principal international models examined in this study. Whereas the European Union structured its artificial intelligence policy through a specific horizontal regulation centered on risk classification and the imposition of obligations directly related to artificial intelligence systems, Brazil developed a more gradual and integrated normative architecture supported by legal instruments originally designed to govern digital transformation in a broader sense.

This difference does not mean that the models lack convergence. On the contrary, international guidelines and the Brazilian experience are becoming increasingly aligned, especially regarding the importance attributed to transparency, human oversight, risk management, and the protection of fundamental rights. The influence of the European AI Act on Bill No. 2,338/2023 demonstrates that the international circulation of regulatory ideas plays an important role in constructing national artificial intelligence policies.

The Brazilian experience nevertheless reveals that simply transplanting foreign models is insufficient to address the particularities of the national constitutional system. The central role assigned to the Federal Constitution, the General Data Protection Law, the ANPD's activities, and the initiatives of the National Council of Justice demonstrates that artificial intelligence governance in Brazil results from interaction among different legal regimes and public institutions, forming a more distributed and polycentric governance model.

This finding reinforces the central hypothesis developed in this study. The Brazilian experience demonstrates that the dichotomy between specific regulation of artificial intelligence and technological neutrality does not necessarily represent an exclusive choice. On the contrary, a hybrid model is emerging in which constitutional principles, general rules protecting rights, and specific risk-management mechanisms coexist in a complementary manner. This evolution provides important support for the proposal of a hybrid regulatory architecture developed in the following chapter, in which the effectiveness of governance derives from integrating general legal foundations with specific instruments of technological oversight.

8 CONCLUSIONS

The expansion of artificial intelligence represents one of the principal technologi-

cal transformations of the twenty-first century, increasing the challenges related to the protection of fundamental rights, economic activity, Public Administration, and the functioning of democratic institutions. In this context, this article examined the debate between two regulatory models: specific regulation of artificial intelligence, represented by the European Union's Artificial Intelligence Act, and the approach based on technological neutrality, according to which the law should regulate the effects of technologies rather than their nature.

The analysis demonstrated that both models have advantages and limitations. Specific regulation facilitates the management of risks related to algorithmic opacity, automated discrimination, and the protection of fundamental rights, but it may rapidly become obsolete in the face of technological evolution. Technological neutrality, in contrast, provides greater normative stability, although it may be insufficient to address particular features of current artificial intelligence systems, such as machine learning and generative models.

The international experience reveals a trend toward convergence around mechanisms based on risk management, transparency, human oversight, and accountability. In Brazil, a governance model has gradually been constructed on the basis of the Federal Constitution, the General Data Protection Law, public policies, and regulatory initiatives such as Bill No. 2,338/2023 and CNJ Resolution No. 615/2025, revealing an integrated normative architecture adapted to the specific features of the national legal system.

It is concluded that the opposition between specific regulation and technological neutrality should not be understood as exclusive. Artificial intelligence governance is more appropriate when structured through a hybrid regulatory model capable of integrating constitutional principles, risk management, transparency, human oversight, and accountability mechanisms. This approach seeks to reconcile technological innovation, legal certainty, and the protection of fundamental rights, offering greater adaptability to the continuous transformations of artificial intelligence and contributing to the development of more effective and sustainable public policies.

REFERENCES

BAROCAS, Solon; SELBST, Andrew D. *Big Data's Disparate Impact*. California Law Review, Berkeley,

v. 104, n. 3, p. 671–732, 2016.

BIONI, Bruno Ricardo. *Proteção de Dados Pessoais: a função e os limites do consentimento*. 3. ed. Rio de Janeiro: Forense, 2023.

BRASIL. Constituição (1988). *Constituição da República Federativa do Brasil de 1988*. Brasília, DF: Presidência da República, 1988.

BRASIL. Lei nº 13.709, de 14 de agosto de 2018. *Lei Geral de Proteção de Dados Pessoais (LGPD)*. Diário Oficial da União: Brasília, DF, 15 ago. 2018.

BRASIL. Lei nº 12.965, de 23 de abril de 2014. *Marco Civil da Internet*. Diário Oficial da União: Brasília, DF, 24 abr. 2014.

BRASIL. Projeto de Lei nº 2.338, de 2023. Dispõe sobre o uso da Inteligência Artificial. Senado Federal, Brasília, DF, 2023.

BRASIL. Ministério da Ciência, Tecnologia e Inovações. *Estratégia Brasileira de Inteligência Artificial (EBIA)*. Brasília: MCTI, 2021.

CONSELHO NACIONAL DE JUSTIÇA. Resolução nº 615, de 11 de março de 2025. Dispõe sobre diretrizes para o desenvolvimento, utilização e governança da inteligência artificial no Poder Judiciário. Brasília: CNJ, 2025.

CRAWFORD, Kate. *Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence*. New Haven: Yale University Press, 2021.

CUSTERS, Bart; FOSCH-VILLARONGA, Eduard (org.). *Law and Artificial Intelligence*. Cham: Springer, 2022.

DIGNUM, Virginia. *Responsible Artificial Intelligence: How to Develop and Use AI in a Responsible Way*. Cham: Springer, 2019.

DONEDA, Danilo. *Da privacidade à proteção de dados pessoais*. 3. ed. São Paulo: Thomson Reuters Brasil, 2021.

ELLUL, Joshua. *Should We Regulate Artificial Intelligence or Some Uses of Software?* In:

CUSTERS, Bart; FOSCH-VILLARONGA, Eduard (org.). *Law and Artificial Intelligence*. Cham: Springer, 2022.

EUROPEAN PARLIAMENT; COUNCIL OF THE EUROPEAN UNION. *Regulation (EU) 2024/1689 of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act)*. Official Journal of the European Union, 2024.

FINOCCHIARO, Giusella. *The Regulation of Artificial Intelligence*. Cheltenham: Edward Elgar Publishing, 2024.

FLORIDI, Luciano. *The Ethics of Artificial Intelligence: Principles, Challenges, and Opportunities*. Oxford: Oxford University Press, 2023.

HILDEBRANDT, Mireille. *Law for Computer Scientists and Other Folk*. Oxford: Oxford University

Press, 2020.

NARAYANAN, Arvind; KAPOOR, Sayash. *AI Snake Oil: What Artificial Intelligence Can Do, What It Can't, and How to Tell the Difference*. Princeton: Princeton University Press, 2024.

NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY. *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*. Gaithersburg: NIST, 2023.

OECD. *OECD AI Principles*. Paris: Organisation for Economic Co-operation and Development, 2019.

PASQUALE, Frank. *The Black Box Society: The Secret Algorithms That Control Money and Information*. Cambridge: Harvard University Press, 2015.

RUSSELL, Stuart; NORVIG, Peter. *Artificial Intelligence: A Modern Approach*. 4. ed. Hoboken: Pearson, 2021.

SCHERTEL MENDES, Laura. *Proteção de Dados, Privacidade e Inteligência Artificial*. São Paulo: Thomson Reuters Brasil, 2022.

UNESCO. *Recommendation on the Ethics of Artificial Intelligence*. Paris: UNESCO, 2021.

WACHTER, Sandra; MITTELSTADT, Brent; FLORIDI, Luciano. Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation. *International Data Privacy Law*, Oxford, v. 7, n. 2, p. 76–99, 2017

FROM AI SYSTEMS TO AUTOMATED ACTIVITIES: FUTURE TRENDS IN ARTIFICIAL INTELLIGENCE REGULATION

Rodrigo da Silva Ferreira

Master in Public Administration and Ph.D. Candidate in Administration at FGV/EBAPE. Attorney and Data Protection Officer (DPO) at the Casa da Moeda do Brasil.

Lecturer and Consultant on Regulation and Emerging Technologies at DecisionLab.

DOI: https://doi.org/10.59224/dti8.ch3

Abstract:. Artificial intelligence regulation has become a central concern for scholars and policymakers. This article argues that AI-centered regulatory frameworks will become increasingly inadequate as automation becomes ubiquitous and embedded in social institutions. Drawing on risk society, sociotechnical systems, and adaptive governance theory, it proposes an expected shift toward activity-based governance supported by a layered regulatory architecture capable of remaining effective under conditions of technological uncertainty and conceptual instability.

Keywords: artificial intelligence regulation; automation governance; risk society.

Table of contents: 1. Introduction; 2. The Limits of AI-Centered Regulation; 2.1. Technological Risk and the Regulatory Problem of Uncertainty; 2.2. AI as a Problematic Regulatory Object; 3. Toward Activity-Based Governance; 3.1. From AI Systems to Sociotechnical Activities; 3.2. A Layered Model for Future Regulation; 3.3. Building Institutional Capacity; 4. Conclusion; References.

1. INTRODUCTION

Writing about future trends for artificial intelligence (AI) regulation is an inherently difficult task. The future is unavailable for direct observation, and the evolution

of the regulatory landscape depends upon complex interactions among technological innovation, institutional choices, economic incentives, geopolitical dynamics, and unforeseen events. The challenge becomes even greater in a risk society[1], a reality of an increasingly *"high technological frontier which absolutely no one completely understands and which generates a diversity of possible futures"*[2]. Under such conditions, developments may unfold in unforeseen directions and at a pace that conventional forecasting fails to anticipate, rendering even apparently well-founded predictions rapidly obsolete (sometimes between a manuscript's submission and its publication [3]).

Those limitations, however, do not make systematic reflection about future trends futile. Foresight derives its value not from predicting the future with precision but from identifying plausible trajectories, anticipating emerging governance challenges, and informing decisions that must be made today. As George Box famously observed, "all models are wrong, but some are useful"[4]. The same epistemic humility should guide any inquiry into future trends in regulation: the scenarios and projections developed through such an exercise will inevitably simplify reality and may ultimately prove inaccurate, but they can nevertheless provide useful frameworks for reasoning and action at the present.

Such reflections are becoming increasingly relevant. Over the past decade, AI has evolved from a relatively specialized area of computer science into a central concern of public law, corporate governance, industrial policy, human rights, national security, labor relations, and geopolitics. Governments have proposed or enacted comprehensive AI-specific legislation[5], regulatory agencies have issued guidance and risk-

---

1 Ulrich Beck, *From Industrial Society to the Risk Society: Questions of Survival, Social Structure and Ecological Enlightenment*, 9 Theory Cult. Soc. 97. (1992).

2 Anthony Giddens, *Risk and Responsibility*, 62 Mod. Law Rev. 1, 3 (1999).

3 For the record, this paper was submitted on July 31, 2026. Readers are invited to assess whether its predictions survived the test of time.

4 *G. E. P. Box, Robustness in the Strategy of Scientific Model Building, in Robustness in Statistics 201, 202 (ROBERT L. Launer & GRAHAM N. Wilkinson eds., 1979), https://www.sciencedirect.com/science/article/pii/B9780124381506500182.*

5 For example, the European Union has enacted the AI Act, while Brazil's National Congress is considering Bill No. 2338/2023, which has already been approved by the Federal Senate.

management frameworks[6], standardization bodies have produced technical norms[7], and organizations have adopted internal governance structures to manage AI risks[8].

At the same time, states have increasingly used regulatory and economic instruments to shape the development of domestic AI capabilities through research funding, public procurement, tax incentives, investment policies, and innovation programs, while also employing export controls, restrictions on advanced semiconductors and computing infrastructure, foreign investment screening, and other measures aimed at preserving technological advantages or limiting strategic competitors [9]. AI governance has therefore become not only a mechanism for mitigating societal risks but also an instrument of industrial strategy and geopolitical competition. The speed and scope of these developments suggest that "AI regulation" has become a broad category through which governments and organizations pursue multiple and sometimes competing objectives, even in scenarios where the precise nature of the underlying technological risks and opportunities remains unclear[10].

This regulatory expansion to control risks while preserving technological advantages has been built over a very unstable conceptual foundation. “Artificial intelligence” does not constitute a single technology or architecture. In fact, even among experts, there is no widely accepted definition of AI[11]. Russell & Norvig propose four

---

6 For example, the French Data Protection Authority (CNIL) has issued recommendations on GDPR compliance in the development of AI systems. See *AI System Development: CNIL's Recommendations to Comply with the GDPR*, available at: https://www.cnil.fr/en/ai-system-development-cnils-recommendations-to-comply-gdpr.

7 For example, the U.S. National Institute of Standards and Technology (NIST) has developed the *Artificial Intelligence Risk Management Framework (AI RMF 1.0)* to guide organizations in identifying, assessing, and managing AI-related risks. Available at: https://www.nist.gov/itl/ai-risk-management-framework.

8 Erich Prem, *From Ethical AI Frameworks to Tools: A Review of Approaches*, 3 AI ETHICS 699 (2023).

9 Kawsar Uddin Mahmud, *From Silicon Valleys to Superpowers: The US–China Tech Rivalry in the Age of Artificial Intelligence*, 189 WORLD AFF., 8 (2026), https://onlinelibrary.wiley.com/doi/abs/10.1002/waf2.70049.

10 Maria Paola Ferretti, *Risk and Distributive Justice: The Case of Regulating New Technologies.*, 16 SCI. ENG. ETHICS 501 (2010).

11 Matthew U. Scherer, *Regulating Artificial Intelligence Systems: Risks, Challenges, Competencies,*

different possible conceptual perspectives: *(i)* systems that act humanly (the Turing test approach), *(ii)* systems that think humanly (the cognitive modeling approach), *(iii)* systems that think rationally (the "laws of thought" approach), and *(iv)* systems that act rationally (the rational agent approach)[12]. Each perspective presents its unique variations, with no indication of convergence. Contemporary regulatory definitions have generally responded by adopting broad functional concepts for "artificial intelligence systems"[13]. This strategy aims to provide some flexibility given the seemingly unsurpassable problem of defining what AI is, but this flexibility comes at a cost. The more fluid the concept of AI system becomes, the harder it is to align regulatory and governance obligations with the concrete activities through which risks actually materialize[14].

This paper argues that the future of AI regulation will not be characterized by continuous refinement of AI-specific definitions but, on the contrary, the gradual displacement of AI itself as the primary object of governance. As highly automated systems become ubiquitous across transportation, healthcare, finance, education, public administration, communication, manufacturing, scientific research, and countless other domains, regulatory attention is likely to shift from AI systems as a technological category toward the governance of automated activities. Such a transition reflects the need to preserve regulatory effectiveness under conditions of rapid technological change and conceptual uncertainty. Ultimately, the future of AI regulation lies not in regulating "AI" itself, but in governing the increasingly automated activities through which risks, responsibilities, and legal consequences arise.

This argument does not suggest that AI-specific regulation is unnecessary at the

---

and Strategies, 29 Harv. J. Law Technol. 353, 359 (2015).

12 For Stuart Russell & Peter Norvig, Artificial Intelligence: A Modern Approach 1–4 (2021).

13 ISO/IEC 22989:2022 defines an AI system as an "engineered system that generates outputs such as content, forecasts, recommendations, or decisions for a given set of human-defined objectives." A closely related definition was adopted by the OECD in its revised *Recommendation on Artificial Intelligence* and subsequently incorporated, with minor adaptations, into the European Union's AI Act.

14 Samuel Carey, *Regulating Uncertainty: Governing General-Purpose AI Models and Systemic Risk*, 17 Eur. J. Risk Regul. 123 (2026).

present. AI governance frameworks may perform important functions by establishing common terminology, baseline governance obligations, and minimum organizational safeguards[15]. However, they also reveal the limits of a technology-centered paradigm. Many of the harms commonly associated with AI do not arise from AI systems in isolation. Discrimination in hiring, unsafe autonomous mobility, erroneous medical triage, manipulative recommender systems, deepfake, undue arrests through biometric identification, and errors in automated public benefits decisions all emerge from sociotechnical activities that combine models, data, hardware, organizational incentives, human decisions, institutional routines, and legal landscapes [16]. Some of these harms may arise from machine learning models or systems that have been conceptualized as artificial intelligence systems. Others may arise from deterministic rules, sensors, actuators, workflow design, data governance failures, or the institutional context in which automation is deployed. Consequently, regulatory systems organized exclusively around AI definitions risk both overinclusion, by imposing obligations on relatively benign computational tools, and underinclusion, by overlooking harmful forms of automation that fall outside prevailing legal definitions of AI systems.

The future of what is currently presented as "AI regulation" will need to address these scope mismatch challenges. It will also need to be consistent with technological evolution and societal trajectories that cannot be fully anticipated[17]. This paper therefore reframes future AI regulation as a problem of institutional preparedness for highly automated societies. Rather than asking which technologies should be regulated, it asks what kind of regulatory architecture can remain legitimate, effective, and protective of fundamental rights as we move into ubiquitous high levels of automation.

To that end, this paper makes three main contributions. First, it argues that the

---

[15] Mattis Jacobs & Judith Simon, *Assigning Obligations in AI Regulation: A Discussion of Two Frameworks Proposed By the European Commission*, 1 DIGIT. SOC. 6 (2022).

[16] Mateusz Dolata, Stefan Feuerriegel & Gerhard Schwabe, *A Sociotechnical View of Algorithmic Fairness*, 32 INF. SYST. J. 754 (2022).

[17] Piret Tõnurist & Angela Hanson, *Anticipatory Innovation Governance: Shaping the Future through Proactive Policy Making*, OECD WORK. PAP. PUBLIC GOV. (2020), https://www.oecd.org/en/publications/anticipatory-innovation-governance_cce14d80-en.html.

most significant long-term trend in AI regulation is the transition from technology-centered regulation toward activity-centered governance, in which automated social activities, rather than AI systems themselves, become the primary unit of regulatory analysis. Second, it integrates insights from risk society theory, sociotechnical systems, and adaptive governance into a theoretical framework explaining why increasingly automated societies require governance models that remain effective despite continuing technological uncertainty. Third, it proposes a layered regulatory architecture combining human rights, constitutional principles, horizontal governance, activity-specific regulation, and adaptive institutional capacity as a more durable foundation for governing future forms of automation.

The remainder of the paper proceeds as follows. Section 2 examines the limitations of AI-centered regulation as a governance paradigm, arguing that technological uncertainty and the conceptual instability of AI increasingly undermine technology-centered approaches. Section 3 develops an alternative framework based on activity-centered governance, proposing a layered regulatory architecture supported by adaptive institutional capacity for increasingly automated societies. The paper concludes by discussing the implications of this paradigm shift for law, public administration, and future research.

2. THE LIMITS OF AI-CENTERED REGULATION

2.1. Technological Risk and the Regulatory Problem of Uncertainty

AI regulation must be understood within a broader transformation in the nature of technological risk. Modern legal systems were largely designed around assumptions of identifiable actors, relatively stable causal chains, and risks that could be managed through ex ante rules, ex post liability, technical standards, or administrative oversight. Emerging technologies challenge each of these assumptions. They generate risks that are increasingly systemic, distributed, adaptive, and only partially observable when regulatory decisions must be made. AI regulation is therefore not merely a specialized branch of technology law, but one manifestation of a broader institutional challenge: governing manufactured risks under conditions of persistent

uncertainty before and while they proliferate[18].

Theories of the risk society - despite many relevant critics[19] - provide a useful framework for understanding this transformation because they shift attention from isolated hazards to the social production and governance of technological risk. Beck argued that contemporary societies increasingly generate hazards through their own systems of production, expertise, and institutional decision-making rather than through external natural events[20]. Giddens similarly described modern technological risks as *manufactured risks*, arising from the very processes intended to increase efficiency, prediction, and control[21]. AI exemplifies this paradox. Automated systems promise greater accuracy, productivity, and personalization while simultaneously introducing new forms of opacity, systemic dependency, and loss of control[22].

These characteristics distinguish AI from many earlier technological domains. First, automated systems can scale decisions at speeds and volumes far beyond ordinary human oversight, allowing errors to affect thousands or even millions of individuals before they are detected[23]. Second, they generate significant informational asymmetries between developers, deployers, regulators, courts, and affected persons, as technical knowledge and system behavior are often unevenly distributed and difficult to interpret[24]. Third, AI systems are frequently embedded within broader sociotechnical infrastructures, making it difficult to isolate the causal contribution of algorithms, datasets, interfaces, organizational procedures, or human intervention to

---

18 Ulrich Kühn & Heather Williams, *Governing the Impact of Emerging Technologies: Actors, Technologies, and Regulation*, 48 J. STRATEG. STUD. 559 (2025).

19 Tom R. Burns & Nora Machado, *Technology, Complexity, and Risk: A Social Systems Perspective on the Discourses and Regulation of the Hazards of Socio-Technical Systems*, SOCIOL. PROBL. E PRÁTICAS (2010), https://revistas.rcaap.pt/sociologiapp/article/view/7803.

20 Beck, *supra* note 1.

21 Giddens, *supra* note 2.

22 John Zerilli et al., *Algorithmic Decision-Making and the Control Problem*, 29 MINDS MACH. 555 (2019).

23 Rebecca Williams, *Accountable Algorithms: Adopting the Public Law Toolbox Outside the Realm of Public Law*, 75 CURR. LEG. PROBL. 237 (2022).

24 Simon Chesterman, *Through a Glass, Darkly: Artificial Intelligence and the Problem of Opacity*, 69 AM. J. COMP. LAW 271 (2021).

a particular outcome[25]. The resulting governance challenge extends beyond the behavior of individual models or systems to the institutional environments in which they operate.

Consequently, the central regulatory problem is not simply one of regulatory lag [26]. Regulatory lag assumes that technological innovation advances faster than legal adaptation, implying that faster legislation or more agile rulemaking would largely mitigate the problem. Although timing certainly matters, it is not the principal challenge. Even perfectly synchronized regulation would continue to face uncertainty regarding future technological capabilities, deployment contexts, dependencies, and harms. Because of their nature, emerging technologies are frequently regulated before their risks are fully understood, requiring institutions to govern not only known hazards but also plausible trajectories of future risk[27]. The regulatory object is therefore unstable not only because technology evolves rapidly, but because its social functions, institutional embedding, and known risk profile evolve alongside its adoption.

This distinction has important implications for the future of AI governance. If uncertainty is structural rather than merely temporal, effective regulation cannot depend exclusively on ever more detailed statutory definitions or on accelerating legislative cycles. Instead, it requires institutions capable of learning, adapting, and revising governance arrangements as technologies evolve and new uncertainties emerge. Regulatory effectiveness should therefore be assessed not only by the substantive obligations imposed at a given moment, but also by the institutional capacity to monitor technological change, integrate new knowledge, coordinate expertise across sectors, update governance mechanisms, and preserve democratic accountability under conditions of continuous technological transformation.

This perspective provides the foundation for the remainder of this paper. The principal challenge for future AI regulation is not merely to govern increasingly sophisticated computational techniques, but to design governance frameworks

---

[25] Dolata, Feuerriegel, and Schwabe, *supra* note 16.

[26] John Bagby & Nizan Packin, *RegTech and Predictive Lawmaking: Closing the RegLag Between Prospective Regulated Activity and Regulation*, 10 MICH. BUS. ENTREP. LAW REV. 127 (2021).

[27] Tõnurist and Hanson, *supra* note 17.

capable of remaining effective despite persistent technological uncertainty. The following sections argue that achieving this objective requires shifting attention from AI as a technological category toward the automated activities through which social risks are ultimately produced.

2.2. AI as a Problematic Regulatory Object

The governance of artificial intelligence is complicated not only by technological uncertainty but also by the conceptual instability of the object being regulated[28], which produces a fundamental problem of scope misalignment: obligations attached to what legal frameworks define as "AI systems" do not necessarily correspond to the activities that generate the harms regulation intends to address.

On the one hand, broad definitions of AI systems may produce overinclusion by capturing low-risk systems merely because they satisfy an abstract conceptualization. Sophisticated models used to optimize internal workflows, organize documents, or support routine operational tasks may trigger costly compliance obligations despite posing minimal risks to fundamental rights, public safety, or democratic institutions. Excessively broad conceptualization may therefore impose disproportionate compliance burdens while diverting regulatory attention away from genuinely high-risk activities.

On the other hand, AI-centered regulation also risks underinclusion by excluding harmful forms of automation that fall outside prevailing definitions of AI systems. Many of the concerns commonly associated with AI, such as discrimination, opacity, surveillance, exclusion, unsafe automation, and procedural unfairness, can emerge from deterministic software, rule-based systems, statistical scoring models, conventional databases, or organizational workflows that incorporate little or no AI systems [29]. In fact, an automated hiring platform based on rigid filtering rules may discrimi-

---

[28] Łukasz Abramowicz, *One Does Not Simply Create an AI Issue. On the Pseudo-Problematic Nature of AI Issue*, 58 Człow. Społecz. 131 (2024).

[29] Mike Zajko, *Automated Government Benefits and Welfare Surveillance*, 21 Surveill. Soc. 246, 248 (2023).

nate unlawfully without using AI in any contemporary technical sense[30]. Likewise, automated credit assessment, welfare administration, or industrial control systems may generate significant legal and societal risks regardless of whether they satisfy statutory definitions of AI[31]. In these cases, the source of risk lies not in artificial intelligence as a technological category, but in the automation of socially significant activities.

The same instability extends beyond definitions of AI itself to the governance concepts surrounding it. Terms such as *fairness*, *transparency*, *accountability*, *explainability*, *robustness*, and *bias* circulate across computer science, law, ethics, public policy, and management, often carrying different meanings in each discipline. Bias may refer either to an ordinary statistical phenomenon or to unlawful discrimination[32]. Likewise, the concepts of explainability and interpretability remain contested[33] with even leading standards organizations, such as ISO and NIST, adopting mutually incompatible conceptualizations[34]. Consequently, governance instruments built upon these concepts may become either technically infeasible or normatively indeterminate if interdisciplinary and conceptual differences remain unaddressed. The challenge is therefore not limited to defining AI, but extends to establishing sufficiently coherent governance vocabularies capable of supporting legal certainty across multiple epistemic communities.

These conceptual difficulties expose the limits of AI-centered regulation as the

---

30 Lindsey Fuchs, *HIRED BY A MACHINE: CAN A NEW YORK CITY LAW ENFORCE ALGORITHMIC FAIRNESS IN HIRING PRACTICES?*, 28 FORDHAM J. CORP. FINANC. LAW 185 (2023).

31 Janine S. Hiller & Lindsay Sain Jones, *Who's Keeping Score?: Oversight of Changing Consumer Credit Infrastructure*, 59 AM. BUS. LAW J. 61 (2022).

32 Jude Browne, Eleanor Drage & Kerry McInerney, *Tech Workers' Perspectives on Ethical Issues in AI Development: Foregrounding Feminist Approaches*, 11 BIG DATA SOC. 20539517231221780 (2024).

33 Matthieu Bellucci et al., *Towards a Terminology for a Fully Contextualized XAI*, 192 PROCEDIA COMPUT. SCI. 241 (2021).

34 ISO/IEC 22989:2022 and the NIST *Artificial Intelligence Risk Management Framework* (AI RMF 1.0) adopt incompatible conceptualizations of explainability and interpretability, effectively inverting the relationship between the two concepts by assigning the "why" of a system's output to one term and the "how" to the other.

primary organizing principle for future AI governance. Comprehensive AI frameworks may perform some valuable functions by establishing common terminology, signaling political priorities, and creating baseline obligations applicable across sectors. Nevertheless, the decisive regulatory questions rarely concern AI as an abstract technological category, and societal expectations will rarely be met with a single perspective for bias or explainability. Instead, the decisive regulatory questions concern the institutional contexts in which automated systems are deployed, the rights and interests affected, the distribution of responsibilities among actors, and the safeguards necessary to prevent or mitigate harm. Whether an automated system supports medical diagnosis, judicial decision-making, financial services, education, public administration, or entertainment and the controls in place is often more consequential from a regulatory perspective than whether it satisfies a particular fixed statutory definition of AI system, bias, or explainability.

The implication is that future regulatory frameworks, if they are to address increasing automation effectively, will need to avoid treating AI as their primary regulatory subject. No definition is likely to remain durable as computational techniques evolve, technological boundaries blur, and capabilities migrate across software, robotics, and other forms of automation. Rather than attempting to resolve conceptual debates about artificial intelligence, regulation should be designed to remain effective despite conceptual instability. This requires shifting the primary unit of regulatory analysis away from technological artifacts and toward the automated activities through which social risks, institutional responsibilities, and legal consequences are ultimately produced. The following section develops this argument by proposing sociotechnical activities, rather than AI systems themselves, as the most appropriate foundation for future regulatory design.

3. TOWARD ACTIVITY-BASED GOVERNANCE

3.1. From AI Systems to Sociotechnical Activities

The conceptual limitations of AI-centered regulation point toward a broader theoretical question: what, precisely, should be the primary object of governance? Responsible AI scholarship has made substantial progress by articulating normative principles such as fairness, accountability, transparency, privacy, safety, robustness,

human oversight, and non-discrimination[35]. These principles have provided a common normative vocabulary through which governments, firms, researchers, and civil society organizations can express concerns about automated decision-making and translate fundamental rights into governance expectations. They have also shaped the design of legislation, technical standards, impact assessments, and corporate governance frameworks across multiple jurisdictions.

Despite their importance, principle-based approaches have well-recognized limitations. Principles are often abstract, potentially conflicting, and insufficiently specified to determine concrete institutional obligations[36]. Fairness may require trade-offs with predictive accuracy[37]. Human oversight may constitute a meaningful safeguard in some contexts while becoming little more than symbolic approval in others[38]. Likewise, explainability may be indispensable where individuals must challenge administrative or judicial decisions, yet comparatively less important in domains where validation, robustness testing, and continuous monitoring provide stronger safeguards. The challenge, therefore, is not whether these principles are desirable, but that they do not themselves specify the institutional mechanisms through which they should be implemented[39].

A sociotechnical perspective addresses this limitation by shifting attention from technological artifacts to the broader systems within which they operate. This suggests that the primary unit of regulatory analysis should not be the AI system itself, but the activity in which it is embedded. An activity is more than a technological use case. It is a socially organized function performed within a particular institutional setting, involving identifiable actors, legal obligations, decision-making procedures,

---

35 Natalia Díaz-Rodríguez et al., *Connecting the Dots in Trustworthy Artificial Intelligence: From AI Principles, Ethics, and Key Requirements to Responsible AI Systems and Regulation*, 99 INF. FUSION 101896 (2023).

36 Malak Sadek et al., *Challenges of Responsible AI in Practice: Scoping Review and Recommended Actions*, 40 AI SOC. 199, 202 (2025).

37 Guilherme Alves et al., *Survey on Fairness Notions and Related Tensions*, 11 EURO J. DECIS. PROCESS. 100033 (2023).

38 *Dolata, Feuerriegel, and Schwabe, supra note 16.*

39 Ayodeji Olusegun Ibitoye, Makuochi Samuel Nkwo & Rita Orji, *Rethinking Responsible AI from Ethical Pillars to Sociotechnical Practice*, 5 AI ETHICS 6207 (2025).

and affected interests. Automated hiring, medical diagnosis, autonomous driving, credit allocation, predictive policing, public benefits administration, algorithmic management, and synthetic media generation are not simply different applications of AI. They are distinct sociotechnical activities, each characterized by its own institutional context, distribution of responsibilities, risk profile, and legal consequences.

This distinction clarifies why identical technologies frequently require different regulatory responses. A large language model used to generate marketing content, summarize internal documents, assist physicians, support judicial reasoning, or influence electoral campaigns may rely on essentially the same computational architecture. Yet the legal significance of these deployments differs profoundly because the underlying activities differ. The relevant regulatory questions concern not only the technical properties of the model but also the rights at stake, the institutional setting, the affected populations, and the consequences of erroneous or abusive decisions. The same technology can therefore justify markedly different governance requirements depending on the activity being automated.

Accordingly, this activity-based governance should begin with four analytical questions. First, (i) what social function is being automated? This identifies whether the relevant activity concerns hiring, lending, healthcare, transportation, policing, education, public administration, or another domain. Second, (ii) who is affected? This requires identifying the individuals, organizations, and communities whose rights, opportunities, or interests may be influenced by the automated activity. Third, (iii) which rights and interests are implicated? Depending on the context, these may include safety, equality, privacy, due process, labor rights, consumer protection, freedom of expression, democratic integrity, or national security. Finally, (iv) which safeguards are institutionally appropriate? The answer may include human review, certification, auditing, transparency obligations, incident reporting, professional standards, licensing, liability rules, procurement restrictions, or, in exceptional circumstances, outright prohibition. These questions provide a more stable and practically relevant foundation for governance than the binary inquiry into whether a system qualifies as AI. In fact, this approach makes the AI system qualification totally irrelevant.

An activity-based perspective also resolves the scope misalignment identified in the previous section. Because it focuses on socially significant functions rather than

technological labels, it encompasses both AI and non-AI components of the same risk-generating activity. Autonomous mobility, for example, depends not only on machine learning models but also on sensors, actuators, deterministic control systems, mapping infrastructures, communication networks, human-machine interfaces, traffic regulation, insurance mechanisms, and liability regimes. Likewise, automated medical diagnosis, financial services, workplace management, and public administration combine computational models with institutional procedures, organizational incentives, professional judgment, and legal constraints. Regulating only the AI component of these activities risks overlooking equally important sources of risk and responsibility.

This shift from technological artifacts to sociotechnical activities does not imply abandoning AI governance. However, such horizontal obligations should be understood as foundational rather than comprehensive. The decisive governance questions arise at the level of concrete activities, where technical capabilities intersect with institutional purposes, legal obligations, and societal consequences. In this sense, technology becomes one variable within a broader regulatory assessment rather than the organizing principle of regulation itself.

The theoretical implication is that future AI governance will be required to move from technology-centered regulation toward activity-centered governance. Instead of asking whether a system falls within a particular definition of AI, regulators will need to ask what social activity is being automated, how automation transforms that activity, which rights and interests are affected, and what institutional safeguards are necessary to preserve legitimacy, accountability, and public trust. This approach preserves technological neutrality without becoming risk-blind, while also reducing dependence on continually evolving definitions of AI. It therefore provides a more stable foundation for governing increasingly heterogeneous forms of automation and is the base for the layered regulatory architecture developed in the following section.

3.2. A Layered Model for Future Regulation

If future AI governance is to be organized around automated activities rather than technological artifacts, regulatory design must also move beyond a single, technolo-

gy-centered framework. No individual regulatory instrument can adequately address the diversity of activities, institutional settings, and risk profiles associated with contemporary AI systems. Horizontal AI legislation, sector-specific regulation, constitutional guarantees, and technical standards each address different dimensions of governance and operate at different levels of abstraction. Rather than viewing these approaches as competing alternatives, they are better understood as complementary components of a layered regulatory architecture capable of combining technological neutrality with context-sensitive risk management.

The first layer consists of constitutional and human rights-based principles that apply irrespective of the technology employed or the activity being automated. Equality, due process, legality, proportionality, privacy, freedom of expression, human dignity, access to justice, and democratic accountability establish the normative boundaries within which all automated systems must operate. These commitments precede technological classification and remain applicable whether decisions are made by human actors, deterministic software, machine learning models, or future forms of computational capacity. Their function is not to regulate AI specifically, but to define the legal values that any regulatory framework must preserve.

The second layer comprises horizontal governance obligations applicable across a broad range of automated systems. Contemporary AI legislation has largely developed at this level, establishing baseline duties concerning documentation, risk management, data governance, cybersecurity, transparency, accountability structures, human oversight, lifecycle monitoring, incident reporting, and model evaluation[40]. These horizontal requirements create a common governance vocabulary, establish minimum expectations for developers and deployers, and promote organizational practices that remain valuable regardless of the sector in which automation occurs. Their principal contribution is not to resolve sector-specific regulatory challenges, but to provide a governance infrastructure upon which more specialized regulatory regimes can build. This is where many contemporary AI regulatory frameworks fall short. In seeking to create comprehensive legal regimes for AI, they assume that a single technology-centered framework can adequately govern the wide variety of activities in which automated systems are deployed. Such an ambition is unlikely to

[40] Carey, *supra* note 14.

succeed. Horizontal governance should therefore be limited to baseline obligations, general prohibitions applicable across sectors, and mechanisms for institutional coordination.

The substantive regulatory requirements should instead be concentrated in the third layer: activity-specific regulation, where legal obligations can be tailored to the risks, institutional contexts, and societal interests associated with particular automated activities. This activity-centered layer constitutes the core of the framework proposed in this paper as an expected future trend. Rather than regulating technologies because they fall within the legal category of AI, it regulates the social activities through which automated systems produce legally relevant effects. Once the relevant activity has been identified, governance obligations can be calibrated to its specific risk profile, institutional setting, and protected legal interests. Automated hiring, medical diagnosis, autonomous mobility, biometric identification, credit allocation, judicial decision support, and automated public administration each raise distinct regulatory concerns despite often relying on similar technological foundations.

Consequently, the safeguards required in one activity may be unnecessary, insufficient, or inappropriate in another. Human review, for example, may be indispensable in judicial decisions yet impractical for autonomous driving, where system validation, safety certification, redundancy, and incident response assume greater importance. Likewise, transparency obligations suitable for consumer-facing recommendation systems may differ substantially from those required in healthcare. Activity-based regulation therefore preserves proportionality by linking governance obligations to the social function being automated rather than to the underlying computational technique. It further permits obligations concerning explainability, fairness, and bias to be calibrated to the characteristics of the regulated activity, rather than relying on the uniform conceptualizations typically embedded in comprehensive AI regulatory frameworks.

This perspective also clarifies the relationship between horizontal and sectoral regulation. Activity-based governance does not replace existing legal regimes; rather, it complements and strengthens them. Healthcare regulators possess expertise concerning clinical validation, patient safety, and professional responsibility. Financial regulators understand prudential supervision, market integrity, and systemic risk. Labor authorities are equipped to address workplace surveillance, algorithmic man-

agement, discrimination, and collective bargaining. Courts remain responsible for due process, legality, evidence, and procedural fairness. Automation modifies how these activities are performed, but it rarely alters the underlying legal interests that justify regulation. In this sense, automated credit scoring remains fundamentally a matter of financial regulation, automated diagnosis remains healthcare regulation, and automated public administration remains administrative law. Automation modifies how these activities are performed, but it does not change the institutional contexts that give them legal significance. Consequently, the regulation of automated activities should remain primarily within the competence of the sectoral regulators responsible for those activities, rather than being transferred to a centralized "AI regulator".

The fourth layer consists of technical and operational standards that translate legal expectations into auditable organizational practices. Technical standards do not determine the normative objectives of regulation. Still, they play a critical role in operationalizing them through testing methodologies, validation procedures, documentation requirements, logging practices, cybersecurity controls, provenance mechanisms, monitoring systems, audit protocols, red-teaming exercises, and performance evaluation. These instruments provide the technical infrastructure necessary to demonstrate compliance and ensure that legal principles can be implemented consistently across organizations and jurisdictions. Their flexibility also allows governance mechanisms to evolve more rapidly than legislation, reducing the need for continual statutory revision as technologies develop.

A layered architecture offers important advantages over purely technology-centered regulation. First, legal obligations remain anchored in enduring legal values and socially significant activities while retaining sufficient flexibility to account for differences in computational methods, deployment contexts, and operational risks at a given moment. Second, it enhances future adaptability. Because regulatory obligations are organized around activities rather than specific technological labels, the framework remains applicable even as technology evolves into increasingly heterogeneous forms of automation. Future systems may integrate machine learning, robotics, neurotechnology, distributed autonomous agents, quantum computing, and entirely new paradigms of human-machine integration that cannot yet be anticipated. A regulatory framework organized exclusively around contemporary defini-

tions of AI systems would require continual conceptual revision. By contrast, a framework centered on automated activities can accommodate technological change without repeatedly redefining its fundamental regulatory categories.

This layered model also addresses two persistent shortcomings of AI governance. The first is normative abstraction without implementation. High-level principles such as fairness, accountability, or transparency remain insufficient unless translated into concrete institutional obligations appropriate to particular activities. The second is technical implementation without normative orientation. Audits, benchmarks, certifications, and documentation practices contribute little unless they are explicitly connected to legal rights, institutional purposes, and democratic values. Layered governance bridges these dimensions by linking constitutional principles to horizontal governance, horizontal governance to activity-specific regulation, and activity-specific regulation to operational standards. Each layer performs a distinct function while reinforcing the others.

The theoretical consequence is that contemporary AI regulation should be viewed as a transitional component of a broader governance framework for increasingly automated societies. The remaining challenge, therefore, is not merely designing appropriate regulatory instruments, but ensuring that public institutions possess the capacity to implement, revise, and enforce them under conditions of continuous technological change. This institutional dimension is the subject of the following section.

3.3. Building Institutional Capacity

The layered architecture proposed above shifts the central challenge of automation governance from defining technological categories to building institutions capable of governing continuously evolving forms of automation[41]. If technological uncertainty, conceptual instability, and sociotechnical complexity are enduring challenges, then regulatory effectiveness depends as much on institutional capacity as on the substantive content of legal rules. Rights, prohibitions, liability regimes, and sector-specific obligations remain indispensable. Still, their effectiveness ultimately

[41] Julie E. Cohen, *The Regulatory State in the Information Age*, 17 THEOR. INQ. LAW 369 (2016).

depends on whether public institutions possess the expertise, resources, and organizational capabilities necessary to interpret, implement, and revise them as technologies evolve[42].

This perspective challenges the traditional assumption that regulation consists primarily of producing legal rules. Automation develops in ways that are iterative, adaptive, and highly context-dependent. New deployment practices, business models, system architectures, and forms of human-machine interaction emerge continuously, often changing the nature of regulatory problems faster than legislation can respond. Under these conditions, regulatory effectiveness depends less on anticipating every future technological development than on establishing institutions capable of learning from technological change.

Rather than treating regulation as a static set of commands, regulators must develop the institutional capacity to sustain adaptive governance through continuous monitoring, experimentation, feedback, evaluation, and organizational learning [43]. Legislative frameworks establish the normative boundaries of acceptable conduct, while complementary governance mechanisms translate those principles into evolving regulatory practice. Regulatory sandboxes, post-market monitoring, incident reporting, conformity assessments, algorithmic audits, certification schemes, impact assessments, technical standards, model evaluations, red-teaming exercises, procurement requirements, and sector-specific guidance exemplify instruments through which regulation can remain responsive without requiring constant legislative revision. Their common purpose is not to replace legislation, but to operationalize legal principles under conditions of persistent uncertainty.

The effectiveness of these mechanisms depends fundamentally on institutional capacity. Regulators will require far more than traditional legal or sectoral expertise. Rather than understanding "AI systems" as a distinct regulatory object, they must develop the technical capacity to understand the forms of automation deployed within the activities they regulate. Depending on the context, this may require expertise in machine learning, robotics, data governance, cybersecurity, human-ma-

---

[42] John W. Bagby & Nizan G. Packin, *Regulatory Technology Shortens Regulatory Lag with Predictive Regulation Targeting Disruptive Technology.*, 41 BANK. FINANC. SERV. POLICY REP. 1 (2022).

[43] Tõnurist and Hanson, *supra* note 17.

chine interaction, model evaluation, or other emerging technologies. Equally important are organizational capacities to process technical documentation, coordinate across regulatory domains, investigate incidents, supervise compliance, and integrate expertise drawn from multiple disciplines. Legal capacity remains essential for translating technical findings into enforceable obligations, while democratic capacity ensures that affected individuals, workers, consumers, firms, and civil society organizations retain meaningful opportunities to challenge decisions that affect their rights and interests[44]. Effective governance therefore depends on the interaction of technical, organizational, legal, and democratic capabilities rather than on any single institutional competency.

Institutional capacity also has an important infrastructural dimension. As automated systems become increasingly complex, effective supervision will require secure data access mechanisms, computational resources, testing environments, interoperable reporting systems, and standardized documentation capable of supporting oversight across multiple sectors. Regulatory agencies that lack such infrastructure may become increasingly dependent upon the technical expertise of the very organizations they regulate, exacerbating existing information asymmetries and potentially undermining regulatory independence. Building institutional capacity therefore requires investment not only in human expertise but also in the technological infrastructure necessary for meaningful supervision.

Paradoxically, maintaining regulatory capacity may itself require increasing reliance on automation. Supervising millions of automated decisions, monitoring complex digital markets, identifying systemic cybersecurity vulnerabilities, or detecting algorithmic harms across distributed infrastructures may exceed the capabilities of purely human oversight. Regulatory agencies are therefore likely to adopt automated tools for risk detection, compliance monitoring, anomaly identification,

---

[44] Nathalie A. Smuha, *The Use of Algorithmic Systems by Public Administrations: Practices, Challenges and Governance Frameworks*, *in* THE CAMBRIDGE HANDBOOK OF THE LAW, ETHICS AND POLICY OF ARTIFICIAL INTELLIGENCE 383 (Nathalie A. Smuha ed., 2025), https://www.cambridge.org/core/books/cambridge-handbook-of-the-law-ethics-and-policy-of-artificial-intelligence/use-of-algorithmic-systems-by-public-administrations/3DA308FE3C0CAF925C08CED7DF0F2697.

inspection prioritization, and evidence analysis[45]. Financial supervision, taxation, cybersecurity, platform governance, and public administration already illustrate this trajectory, where regulators increasingly employ computational systems to oversee computational systems[46].

This evolution gives rise to a profound democratic paradox. Automation may be essential for preserving the effectiveness of public oversight in technologically complex societies, yet automation within regulatory institutions creates new risks for transparency, accountability, contestability, and due process[47]. Public authorities that rely on algorithmic systems to allocate enforcement resources, prioritize inspections, identify regulatory violations, or support administrative decision-making inevitably become subject to the same governance concerns they impose upon private actors. The legitimacy of automated regulation therefore depends not only on its technical performance but also on its conformity with constitutional principles such as legality, proportionality, procedural fairness, explainability, judicial review, and public justification.

Institutional capacity must therefore be understood as both a technical and a constitutional concept. A capable regulator is not simply one equipped with advanced computational tools or specialized expertise. Rather, it is an institution capable of combining technical competence with democratic legitimacy, one that can explain its decisions, justify its methods, revise its practices in light of new evidence, and remain accountable to the legal order it serves. Speed, automation, and analytical sophistication are valuable only insofar as they reinforce, rather than undermine, the rule of law and public trust.

Taken together, these considerations suggest that future regulators will depend less on identifying better definitions for the automated systems and more on designing institutions capable of governing increasingly automated societies under conditions of persistent uncertainty. Conceptual instability favors activity-based regula-

---

45 Karen Yeung, *Algorithmic Regulation: A Critical Interrogation*, 12 Regul. Gov. 505 (2018).

46 Oliver Hackney & Anna Huggins, *Automating Sanctions Corequirempliance: Aligning Regulatory Technology, Rules and Goals*, 5 Law Technol. Hum. 165 (2023).

47 Madalina Busuioc, *Accountable Artificial Intelligence: Holding Algorithms to Account*, 81 Public Adm. Rev. 825 (2021),.

tion over technology-centered definitions, and sociotechnical complexity calls for layered governance that combines constitutional principles, horizontal obligations, activity-specific safeguards, and evolving technical standards. None of these elements, however, can operate effectively without institutions capable of continuous learning, interdisciplinary coordination, and democratic accountability. Regulators should therefore be evaluated by their capacity to remain legitimate, effective, and adaptable as technologies, institutions, and societal expectations continue to evolve.

4. CONCLUSION

This paper has argued that the future of AI regulation is unlikely to be determined by increasingly sophisticated legal definitions of AI systems or horizontal AI frameworks. Although definitional work remains important for legal certainty, institutional coordination, and compliance, it cannot resolve the more fundamental challenge that AI is an evolving, heterogeneous, and ultimately unstable regulatory category. As automation becomes more pervasive across economic sectors and public institutions, the legal significance of what are considered "AI risks" derives less from the technology itself than from the activities through which it reshapes the exercise of power, the allocation of opportunities, and the distribution of risks.

The argument is therefore that "AI regulation" should be understood as a transitional stage in the broader evolution of governance for increasingly automated societies. The critical regulatory question is progressively shifting from *what technology is being used* to *what socially relevant activity is being automated*. This shift recognizes that many of the harms commonly attributed to AI, including discrimination, safety failures, opacity, surveillance, and procedural unfairness, emerge from sociotechnical activities involving computational models, organizational incentives, institutional routines, data infrastructures, and human decision-makers rather than from AI systems considered in isolation. Activity-based governance therefore offers a more stable and technologically neutral foundation for future regulation than frameworks centered exclusively on AI definitions.

Building upon this premise, the paper proposed a layered governance architecture combining human rights and constitutional principles, horizontal governance obligations, activity-specific regulation, and adaptive technical standards. Such an ap-

proach preserves the advantages of horizontal "AI governance" current initiatives while avoiding the overinclusion and underinclusion problems that increasingly accompany technology-centered regulation. It also acknowledges that governance requirements cannot be determined solely by the presence of AI. Still, it must instead reflect the rights affected, the institutional context, and the safeguards appropriate to each automated activity.

The analysis further suggests that the long-term effectiveness of regulation depends less on the quantity or breadth of legal rules than on the capacity of public institutions to implement, adapt, and enforce them. In environments characterized by continuous technological change, regulatory capacity becomes a strategic resource. Institutions capable of monitoring emerging technologies, integrating multidisciplinary expertise, conducting audits, coordinating across regulatory domains, learning from implementation, and adapting governance mechanisms may ultimately prove more important than the substantive legal obligations they enforce.

This institutional perspective also broadens the understanding of AI and automation governance. Contemporary regulatory frameworks pursue objectives that extend far beyond the mitigation of technological risks. AI regulation has simultaneously become an instrument of industrial policy, technological sovereignty, economic competitiveness, national security, and geopolitical strategy. Governments increasingly rely on research funding, public procurement, investment incentives, export controls, semiconductor policies, standards-setting initiatives, and international partnerships not merely to regulate AI, but to shape the global distribution of technological capabilities and strategic advantage. Future governance will therefore require balancing not only innovation and the protection of fundamental rights, but also economic resilience, strategic competition, and international cooperation. Recognizing these multiple and sometimes competing objectives reveals that regulators are no longer concerned solely with mitigating technological risks, but increasingly with shaping the institutional, economic, and geopolitical conditions of an increasingly automated society.

The argument advanced here is necessarily prospective. It does not seek to predict particular technological breakthroughs or propose a definitive taxonomy of future regulatory problems. Rather, it offers a structural claim: as automation becomes increasingly embedded in social, economic, and governmental institutions, regulato-

ry frameworks organized primarily around the concept of AI will become progressively less capable of capturing the activities through which legally significant risks emerge. Future research should therefore explore how activity-based governance can be operationalized across different regulatory domains, how public institutions can strengthen both their technical and democratic capacities, and how governance systems themselves should evolve as regulators increasingly rely on automation.

Ultimately, the future of AI regulation lies not in governing AI as a technological category, but in governing the increasingly automated activities through which technology reshapes economic production, public administration, scientific research, communication, and the exercise of power. The enduring question is not whether regulators can continue organizing the governance of highly automated activities around the legal construct of "AI systems" (an approach that appears increasingly unsustainable for the reasons developed in this paper), but whether democratic institutions can govern the transition toward highly automated societies while preserving human agency, accountability, and the rule of law.

REFERENCES

Abramowicz, Ł. (2024). One does not simply create an AI issue. On the pseudo-problematic nature of AI issue. *Człowiek i Społeczeństwo*, *58*, 131–150. https://doi.org/10.14746/cis.2024.58.7

Alves, G., Bernier, F., Couceiro, M., Makhlouf, K., Palamidessi, C., & Zhioua, S. (2023). Survey on fairness notions and related tensions. *EURO Journal on Decision Processes*, *11*, 100033. https://doi.org/10.1016/j.ejdp.2023.100033

Bagby, J., & Packin, N. (2021). RegTech and Predictive Lawmaking: Closing the RegLag Between Prospective Regulated Activity and Regulation. *Michigan Business & Entrepreneurial Law Review*, *10*(2), 127–178. https://doi.org/https://doi.org/10.36639/mbelr.10.2.regtech

Bagby, J. W., & Packin, N. G. (2022). Regulatory Technology Shortens Regulatory Lag with Predictive Regulation Targeting Disruptive Technology. *Banking & Financial Services Policy Report*, *41*(1), 1–14. (155305772).

Beck, U. (1992). From Industrial Society to the Risk Society: Questions of Survival, Social Structure and Ecological Enlightenment. *Theory, Culture & Society*, *9*(1), 97–123. https://doi.org/10.1177/026327692009001006

Bellucci, M., Delestre, N., Malandain, N., & Zanni-Merk, C. (2021). Towards a terminology for a fully contextualized XAI. *Procedia Computer Science, Knowledge-Based and Intelligent Information & Engineering Systems: Proceedings of the 25th International Conference KES2021*, *192*, 241–250. https://doi.org/10.1016/j.procs.2021.08.025

Box, G. E. P. (1979). Robustness in the Strategy of Scientific Model Building. In R. L. Launer & G. N. Wilkinson (Eds.), *Robustness in Statistics* (pp. 201–236). Academic Press. https://doi.org/10.1016/B978-0-12-438150-6.50018-2

Browne, J., Drage, E., & McInerney, K. (2024). Tech workers' perspectives on ethical issues in AI development: Foregrounding feminist approaches. *Big Data & Society*, *11*(1), 20539517231221780. https://doi.org/10.1177/20539517231221780

Burns, T. R., & Machado, N. (2010). Technology, complexity, and risk: A social systems perspective on the discourses and regulation of the hazards of socio-technical systems. *Sociologia, Problemas e Práticas*, (62). https://revistas.rcaap.pt/sociologiapp/article/view/7803

Busuioc, M. (2021). Accountable Artificial Intelligence: Holding Algorithms to Account. *Public Administration Review*, *81*(5), 825–836. https://doi.org/10.1111/puar.13293

Carey, S. (2026). Regulating Uncertainty: Governing General-Purpose AI Models and Systemic Risk. *European Journal of Risk Regulation*, *17*(1), 123–139. https://doi.org/10.1017/err.2025.10040

Chesterman, S. (2021). Through a Glass, Darkly: Artificial Intelligence and the Problem of Opacity. *The American Journal of Comparative Law*, *69*(2), 271–294. https://doi.org/10.1093/ajcl/avab012

Cohen, J. E. (2016). The Regulatory State in the Information Age. *Theoretical Inquiries in Law*, *17*(2), 369–414. https://doi.org/10.1515/til-2016-0015

Díaz-Rodríguez, N., Del Ser, J., Coeckelbergh, M., López de Prado, M., Herrera-Viedma, E., & Herrera, F. (2023). Connecting the dots in trustworthy Artificial Intelligence: From AI principles, ethics, and key requirements to responsible AI systems and regulation. *Information Fusion*, *99*, 101896. https://doi.org/10.1016/j.inffus.2023.101896

Dolata, M., Feuerriegel, S., & Schwabe, G. (2022). A sociotechnical view of algorithmic fairness. *Information Systems Journal*, *32*(4), 754–818. https://doi.org/10.1111/isj.12370

Ferretti, M. P. (2010). Risk and Distributive Justice: The Case of Regulating New Technologies. *Science & Engineering Ethics*, *16*(3), 501–515. (52799329). https://doi.org/10.1007/s11948-009-9172-z

Fuchs, L. (2023). HIRED BY A MACHINE: CAN A NEW YORK CITY LAW ENFORCE ALGORITHMIC FAIRNESS IN HIRING PRACTICES? *Fordham Journal of Corporate & Financial Law*, *28*(1), 185–222. (161802934).

Giddens, A. (1999). Risk and Responsibility. *The Modern Law Review*, *62*(1), 1–10. https://doi.org/10.1111/1468-2230.00188

Hackney, O., & Huggins, A. (2023). Automating Sanctions Compliance: Aligning Regulatory Technology, Rules and Goals. *Law, Technology and Humans*, *5*(2), 165–180. https://doi.org/10.5204/lthj.2506

Hiller, J. S., & Jones, L. S. (2022). Who's Keeping Score?: Oversight of Changing Consumer Credit Infrastructure. *American Business Law Journal*, *59*(1), 61–121. https://doi.org/10.1111/ablj.12199

Ibitoye, A. O., Nkwo, M. S., & Orji, R. (2025). Rethinking responsible AI from ethical pillars to socio-

technical practice. *AI and Ethics*, *5*(6), 6207–6223. https://doi.org/10.1007/s43681-025-00809-2

Jacobs, M., & Simon, J. (2022). Assigning Obligations in AI Regulation: A Discussion of Two Frameworks Proposed By the European Commission. *Digital Society*, *1*(1), 6. https://doi.org/10.1007/s44206-022-00009-z

Kühn, U., & Williams, H. (2025). Governing the impact of emerging technologies: Actors, technologies, and regulation. *Journal of Strategic Studies*, *48*(3), 559–570. https://doi.org/10.1080/01402390.2025.2472213

Mahmud, K. U. (2026). From Silicon Valleys to Superpowers: The US–China Tech Rivalry in the Age of Artificial Intelligence. *World Affairs*, *189*(1). https://doi.org/10.1002/waf2.70049

Prem, E. (2023). From ethical AI frameworks to tools: A review of approaches. *AI and Ethics*, *3*(3), 699–716. https://doi.org/10.1007/s43681-023-00258-9

Russell, S., & Norvig, P. (2021). *Artificial Intelligence: A Modern Approach*. Pearson.

Sadek, M., Kallina, E., Bohné, T., Mougenot, C., Calvo, R. A., & Cave, S. (2025). Challenges of responsible AI in practice: Scoping review and recommended actions. *AI & SOCIETY*, *40*(1), 199–215. https://doi.org/10.1007/s00146-024-01880-9

Scherer, M. U. (2015). Regulating Artificial Intelligence Systems: Risks, Challenges, Competencies, and Strategies. *Harvard Journal of Law & Technology*, *29*(2), 353–400.

Smuha, N. A. (2025). The Use of Algorithmic Systems by Public Administrations: Practices, Challenges and Governance Frameworks. In N. A. Smuha (Ed.), *The Cambridge Handbook of the Law, Ethics and Policy of Artificial Intelligence* (pp. 383–410). Cambridge University Press. https://doi.org/10.1017/9781009367783.023

Tõnurist, P., & Hanson, A. (2020). Anticipatory innovation governance: Shaping the future through proactive policy making. *OECD Working Papers on Public Governance*. https://doi.org/10.1787/cce14d80-en

Williams, R. (2022). Accountable Algorithms: Adopting the Public Law Toolbox Outside the Realm of Public Law. *Current Legal Problems*, *75*(1), 237–263. https://doi.org/10.1093/clp/cuac008

Yeung, K. (2018). Algorithmic Regulation: A Critical Interrogation. *Regulation & Governance*, *12*(4), 505–523. https://doi.org/10.1111/rego.12158

Zajko, M. (2023). Automated Government Benefits and Welfare Surveillance. *Surveillance & Society*, *21*(3), 246–258. https://doi.org/10.24908/ss.v21i3.16107

Zerilli, J., Knott, A., Maclaurin, J., & Gavaghan, C. (2019). Algorithmic Decision-Making and the Control Problem. *Minds and Machines*, *29*(4), 555–578. https://doi.org/10.1007/s11023-019-09513-7

II
INTERNATIONAL GOVERNANCE AND REGULATION OF ARTIFICIAL INTELLIGENCE

BEYOND THE CLOUD ACT: DIGITAL SOVEREIGNTY, TECHNOLOGICAL HEGEMONY AND INSTITUTIONAL PATHWAYS FOR REGULATION OF ARTIFICIAL INTELLIGENCE IN BRAZIL

Cristiano de Souza Therrien

Lecturer at the Université de Montréal Faculty of Law and Toronto Metropolitan University; Associated Researcher, Centre for Law, Technology and Society (AI + Society Initiative), University of Ottawa. LL.D., Université de Montréal. E-mail: cristiano.therrien@umontreal.ca

Felipe Cittolin Abal

Professor at the Law School and at the Graduate Program in Law (PPGD), Universidade de Passo Fundo (UPF) and History (PPGH), Universidade de Passo Fundo (UPF). Ph.D. in History, Universidade de Passo Fundo.

E-mail: felipeabal@upf.br

DOI: https://doi.org/10.59224/dti8.ch4

Abstract: The digitalization of Brazilian public services has deepened dependence on cloud and compute infrastructure controlled by U.S. corporations, exposing the country to laws asserting extraterritorial jurisdiction over data, notably the CLOUD Act and FISA Section 702. Pairing digital sovereignty and technological hegemony as lenses, this chapter examines that legal architecture, its incompatibilities with the LGPD, and three Brazilian episodes, and proposes a four-axis counter-hegemonic agenda: a sectoral blocking statute, sovereign cloud expansion, AI bill amendments, a bilateral track.

Keywords: Digital Sovereignty; CLOUD Act; Artificial Intelligence Regulation.

Table of contents: Introduction. 1. Digital sovereignty and technological hegemony: two concepts in tension. 2. The CLOUD Act, FISA Section 702 and the legal architecture of extraterritorial access. 3. Brazilian digital dependence: manifestations, risks and institutional responses. 3.1 The São Paulo Court of Justice and Microsoft. 3.2 DATASUS and the health data of the Brazilian population. 3.3 From latent risk to open confrontation: the 2025–2026 sanctions and trade offensive. 3.4 Institutional responses: the sovereign cloud, the AI bill and the data centre policy. 3.5 Why compute, not just data. Conclusion.

INTRODUCTION

The digitalization of Brazilian public services has unfolded through a paradox: the more the state invests in modernization, the more dependent it becomes on digital infrastructure controlled by foreign corporations. The problem is not that these corporations are foreign, which is common enough, but that infrastructural dependence is combined with foreign legislation asserting extraterritorial jurisdiction over data controlled by companies headquartered abroad. The paradigmatic instrument of this assertion is the Clarifying Lawful Overseas Use of Data Act (CLOUD Act), enacted by the United States Congress on 23 March 2018.[1] The question this chapter asks is direct: how does the CLOUD Act, together with the surveillance architecture that surrounds it, give effect to U.S. technological hegemony over Brazil, and what institutional pathways remain open, now that artificial intelligence (AI) has raised the stakes, to reduce the resulting vulnerability?

The CLOUD Act authorizes U.S. authorities to compel the disclosure of data stored by U.S. service providers regardless of where that data is physically located. Even if the data remains in a data centre in Brazil, if the controlling company is American, U.S. authorities may gain access to it, in disregard of Brazilian law.

Brazilian vulnerability has a concrete dimension. As of the first quarter of 2026,

[1] UNITED STATES. *Clarifying Lawful Overseas Use of Data Act (CLOUD Act)*. 18 U.S.C. § 2713. Washington, DC: U.S. Congress, 2018.

three U.S. companies (Amazon Web Services, Microsoft Azure and Google Cloud Platform) jointly held approximately 63% of the global cloud infrastructure market. [2] When Brazilian government data related to health, social security or judicial proceedings is entrusted to these companies, it becomes potentially subject to U.S. authorities upon a CLOUD Act request, contractual safeguards notwithstanding. Moreover, such contracts have at times been concluded without even basic safeguards, leaving data exposed to cyberattacks.[3]

Recent events have moved this vulnerability from the realm of doctrinal concern to that of open geopolitical confrontation. Between 2025 and 2026, the U.S. imposed financial sanctions on a sitting Justice of the Brazilian Supreme Federal Tribunal (STF) under the Global Magnitsky framework,[4] tied steep additional tariffs on Brazilian goods to the conduct of Brazilian courts towards U.S. technology platforms, and directed a Section 301 investigation covering, among other matters, Brazil's digital trade policies and its instant payment system, Pix.[5] These episodes, examined in section 3, confirm that the legal architecture analyzed here is not a dormant reserve of power: it is an operative tool of technological hegemony.

This chapter investigates how the CLOUD Act gives concrete effect to U.S. technological hegemony over Brazil and proposes feasible institutional pathways to reduce this vulnerability. Our central hypothesis is that Brazilian dependence does not result merely from technological deficit or absent national capabilities, but from

---

[2] SYNERGY RESEARCH GROUP. Cloud market annual revenue run rate topped half a trillion dollars in Q1 as growth surge continues. Reno, May 2026. Disponível em: https://www.srgresearch.com/articles/cloud-market-annual-revenue-run-rate-topped-half-a-trillion-dollars-in-q1-as-growth-surge-continues. Acesso em: 1 ago. 2026.

[3] See section 3.2 below.

[4] CONSULTOR JURÍDICO. Lei Magnitsky projeta ordem jurídica dos EUA em território brasileiro. São Paulo, 31 out. 2025. Disponível em: https://www.conjur.com.br/2025-out-31/lei-magnitsky-projeta-ordem-juridica-dos-eua-em-territorio-brasileiro/. Acesso em: 31 jul. 2026.

[5] UNITED STATES. Office of the United States Trade Representative (USTR). Initiation of Section 301 Investigation: Brazil's Acts, Policies, and Practices Related to Digital Trade and Electronic Payment Services. *Federal Register*, Washington, DC, 18 July 2025 (investigation directed on 15 July 2025). Disponível em: https://www.federalregister.gov/documents/2025/07/18/2025-13498/initiation-of-section-301-investigation-brazils-acts-policies-and-practices-related-to-digital-trade. Acesso em: 31 jul. 2026.

public policy choices that naturalize and perpetuate foreign dependence, creating obstacles to the construction of national digital sovereignty. That dependence is path-dependent rather than inevitable: it cannot be quickly or completely reversed, but it can be reduced incrementally for critical public-sector workloads through procurement, technical controls and domestic capability investment.

The analysis rests on three theoretical frameworks. The first is the concept of surveillance capitalism developed by Shoshana Zuboff to describe a new form of accumulation based on the extraction of behavioural surplus and, most relevant here, the close relationship between big tech corporations and the U.S. government.[6] The second is digital colonialism, as described by Couldry and Mejias, for whom data extraction carried out by Global North corporations constitutes a contemporary form of appropriation.[7] The third is the conceptual pair that gives this chapter its title, digital sovereignty and technological hegemony, whose meanings and structural relationship are specified in section 1.

We adopt a qualitative, interdisciplinary approach combining international law, digital law and international relations. We analyze normative texts (the CLOUD Act, the LGPD, the GDPR, Bill No. 2,338/2023), review the literature on digital sovereignty, technological hegemony, digital colonialism and surveillance capitalism, and examine concrete cases, the São Paulo Court of Justice and DATASUS, that evidence the risks identified. In addition, we analyze official documents such as the records of the French Senate inquiry committee hearing with Microsoft France executives (June 2025). This French case matters for a study centred on Brazil and the United States because it supplies what the Brazilian record does not yet contain: a sworn corporate admission that contractual safeguards yield to the CLOUD Act. A note on sources:

---

6 ZUBOFF, Shoshana. *The age of surveillance capitalism: the fight for a human future at the new frontier of power*. New York: PublicAffairs, 2019. For a political-economy corrective emphasizing structural rent extraction rather than behavioural aberration, see RIKAP, Cecilia. Amazon: a story of accumulation through intellectual rentiership and predation. *Competition & Change*, v. 26, n. 3-4, p. 436-466, 2022. DOI: 10.1177/1024529420932418; and the critical review by COHEN, Julie E. Review of Zuboff's The Age of Surveillance Capitalism. *Surveillance & Society*, v. 17, n. 1/2, p. 240-245, 2019.

7 COULDRY, Nick; MEJIAS, Ulises A. *The costs of connection: how data is colonizing human life and appropriating it for capitalism*. Stanford: Stanford University Press, 2019.

for recent Brazilian institutional facts we rely wherever possible on official records (Câmara, Senado, CNJ, ANPD, Serpro), complemented by specialized press where no primary record is public.

The relevance of this work lies in the articulation of the concepts employed, in the analysis of the CLOUD Act and in the formulation of concrete proposals to mitigate national vulnerability in a context of open dispute over global technological hegemony, in which AI has raised the stakes: whoever controls compute, data and models will shape the infrastructures of decision-making itself. The chapter is structured in three sections. The first specifies the chapter's two organizing concepts and their relationship, and situates Brazil's subordinate insertion in the global informational economy. The second analyzes the legal architecture of the CLOUD Act and of Section 702 of the Foreign Intelligence Surveillance Act, their incompatibilities with Brazilian data protection law, and the French Senate hearing in which that exposure was acknowledged. The third examines concrete manifestations of Brazilian digital dependence (the São Paulo Court of Justice and DATASUS cases, and the 2025–2026 sanctions and trade offensive) before assessing the institutional responses under construction: the sovereign cloud, the AI bill (PL 2,338/2023), the data centre policy (REDATA) and the Brazilian AI Plan, with a brief comparative note on Canada. The conclusion presents a four-axis agenda which, we argue, is best understood as a counter-hegemonic strategy.

1 DIGITAL SOVEREIGNTY AND TECHNOLOGICAL HEGEMONY: TWO CONCEPTS IN TENSION

Digital sovereignty has emerged as a key concept in contemporary debates on internet governance and the technological autonomy of states. The term, however, is not a univocal category: it harbours multiple possible meanings depending on the contexts and actors that mobilize it.[8] This plurality of understandings is not a matter

[8] For a systematic synthesis of this plurality, based on a corpus of 271 peer-reviewed articles and distinguishing rights-based, market-oriented, centralisation and state-based models, see FRATINI, Samuele; HINE, Emmie; NOVELLI, Claudio; ROBERTS, Huw; FLORIDI, Luciano. Digital sovereignty: a descriptive analysis and a critical evaluation of existing models. *Digital Society*, v. 3, n. 3, art. 59, 2024. DOI: 10.1007/s44206-024-00146-7.

of terminological imprecision; it reflects ongoing disputes over the role of the state, the market and citizens with respect to digital infrastructures. As Pohle and Thiel observe, digital sovereignty operates less as a settled legal or organizational concept than as a discursive practice in politics and policy: a claim about who ought to control the digital, advanced by actors with divergent projects.[9] In a later critical assessment, Pohle, Nanni and Santaniello suggest that digital sovereignty is best understood *"primarily as a process rather than as a status"*: a continuing negotiation of control over infrastructures, data and standards, rather than a condition a state either possesses or lacks.[10]

Because this chapter pairs two concepts that are related but not synonymous, their use here must be stipulated at the outset. By digital sovereignty we mean a claim about control: the asserted capacity of a state (or, in other formulations, of a collectivity) to govern the data, infrastructures and technological decisions that affect the rights of its population and its strategic interests. Following the discursive-practice school, we treat it as a process and a claim rather than a status.

By technological hegemony we mean a claim about power. Following the Gramscian tradition as brought into international relations by Robert Cox, hegemony is a structure of dominance that rests on consent as much as coercion; in its technological form, the dominant actor's control of the material substrate of digital life (infrastructures, standards, protocols and, increasingly, compute) makes dependence appear natural, technical and freely chosen.[11] Technology is not external to this structure but

---

9 POHLE, Julia; THIEL, Thorsten. *Digital sovereignty. Internet Policy Review*, v. 9, n. 4, 2020. DOI: 10.14763/2020.4.1532.

10 POHLE, Julia; NANNI, Riccardo; SANTANIELLO, Mauro. *Unthinking digital sovereignty: a critical reflection on origins, objectives, and practices*. *Policy & Internet*, v. 16, n. 4, p. 666-671, 2024. DOI: 10.1002/poi3.437.

11 COX, Robert W. Gramsci, hegemony and international relations: an essay in method. *Millennium: Journal of International Studies*, v. 12, n. 2, p. 162-175, 1983. DOI: 10.1177/03058298830120020701 (reprinted in COX, Robert W.; SINCLAIR, Timothy J. *Approaches to world order*. Cambridge: Cambridge University Press, 1996. p. 124-143). The neo-Gramscian reading of hegemony in international relations is contested: see GERMAIN, Randall D.; KENNY, Michael. Engaging Gramsci: international relations theory and the new Gramscians. *Review of International Studies*, v. 24, n. 1, p. 3-21, 1998; and, in reply, IVES, Peter; SHORT, Nicola. On Gramsci and the international: a textual analysis. *Review of International Studies*, v.

its material carrier. As McCarthy argues in the Gramscian register, *"technological structures are the product of particular cultural values and embed these cultural values within their very structure"*, so that physical infrastructures express and stabilize the ideational order that produced them.[12] We use technological hegemony as the broader term, extending beyond data flows to cloud, compute, chips and standards; digital hegemony, as the expression is commonly used in the literature, is its informational face. The two title concepts are structurally linked: sovereignty is the vocabulary in which contested hegemony becomes visible, a relationship developed at the end of this section.

The intuition behind this pairing did not originate in the academic literature alone. It was put to one of the authors in a public debate by the Brazilian philosopher Fabrício Pontin, for whom the present dispute concerns digital sovereignty on one side but turns, on the other, on the more consequential concept of digital hegemony: if information is the great wealth of our time, power lies with whoever holds it.[13] What follows can be read as a legal-institutional elaboration of that insight: sovereignty names the claim Brazil is making; hegemony names the structure that claim runs up against.

Within the sovereignty debate, Couture and Toupin identify at least five categories mobilized by different actors: cyberspace sovereignty, centred on state control of digital communications; governmental sovereignty, directed at the protection of data and essential infrastructures; Indigenous sovereignty, which claims the right of self-determination of Indigenous peoples; the sovereignty of social movements, linked to

39, n. 3, p. 621-642, 2013.

[12] McCARTHY, Daniel R. The meaning of materiality: reconsidering the materialism of Gramscian IR. *Review of International Studies*, v. 37, n. 3, p. 1215-1234, 2011. DOI: 10.1017/S0260210510000077X, p. 1215. See also McCARTHY, Daniel R. Power, information technology, and international relations theory. Basingstoke: Palgrave Macmillan, 2015; and, for the cognate concept of structural power over the knowledge structure, STRANGE, Susan. *States and markets*. London: Pinter, 1988.

[13] PONTIN, Fabrício. [Participação]. In: ABAL, Felipe Cittolin; DIVAN, Gabriel. #471 "A IA vai curar o câncer?". Viracasacas [Podcast], 17 fev. 2026. Disponível em: https://youtu.be/F9MlIgT-kodE. Acesso em: 1 ago. 2026. Trecho a 1h06min37s. The authors acknowledge this provocation as one of the motivations for the present chapter.

technological autonomy and free software; and personal sovereignty, related to control over personal data.[14] Each of these categories carries a different political project, and the projects often contradict one another. This chapter deliberately adopts a state-centric scope, focused on the Brazilian state's capacity to protect public-sector data and infrastructure. The collective and non-state loci of digital sovereignty, Indigenous data sovereignty foremost among them, raise distinct questions beyond the scope of this study.[15]

In the Brazilian context, the ambiguity of the concept deepens when different theoretical traditions are considered. Massuchin, Tavares and Comel distinguish between digital, informational, cyber and data sovereignty. Digital sovereignty in a strict sense concerns the state's capacity to regulate and control infrastructures, platforms and data flows within its territory. Informational sovereignty broadens the scope to include the production, circulation and appropriation of knowledge. Cyber sovereignty centres on the protection of critical systems, while data sovereignty focuses on the legal guarantees afforded to the processing of personal and collective information.[16] Of these senses, we work chiefly with digital sovereignty in the strict sense, joined to the data-sovereignty dimension, as stipulated above.

The attribution of multiple meanings to the term is bound up with current geopolitical tensions and with the recognition that digital technologies, like any technology, are not neutral, but instruments of political and economic power. The Brazilian debate on digital sovereignty gained salience through concrete cases, such as the 2019 contract between Microsoft and the São Paulo Court of Justice, which laid bare the institutional vulnerability of public bodies before foreign suppliers subject to foreign law. Such situations show that the question exceeds technical considerations and

---

[14] COUTURE, Stéphane; TOUPIN, Sophie. What does the notion of "sovereignty" mean when referring to the digital? *New Media & Society*, v. 21, n. 10, p. 2305-2322, 2019.

[15] On Indigenous data sovereignty, see CARROLL, Stephanie Russo et al. The CARE Principles for Indigenous data governance. *Data Science Journal*, v. 19, art. 43, p. 1-12, 2020. DOI: 10.5334/dsj-2020-043.

[16] MASSUCHIN, Michele Goulart; TAVARES, Camila Quesada; COMEL, Naiza. Soberania digital e informacional: desafios da gestão de tecnologias, da comunicação pública e os reflexos para a cidadania. In: ENCONTRO ANUAL DA COMPÓS, 34., 2025, Curitiba. Anais [...]. Curitiba: UFPR, 2025.

involves complex legal, economic and geopolitical dimensions.[17]

This diversity also reveals disputes over the very meaning of sovereignty in the digital context. If state sovereignty traditionally denotes exclusive control over a given territory, networks challenge this logic through transnational data flows that escape physical borders. The tension is acute in Global South countries, where technological dependence compounds historical asymmetries of power. In Brazil's case, insertion into the global digital economy occurs in a subordinate fashion, visible in the concentration of services on foreign platforms and in the country's limited industrial capacity for hardware production and the development of autonomous infrastructures.[18]

The aspiration to digital sovereignty faces obstacles that are not contingent but constitutive of the contemporary digital order, resulting in what may be termed digital colonialism.[19] Brazil's technological dependence manifests itself in several dimensions. As regards infrastructure, Brazil does not produce components such as processors, servers and chips. According to data presented by Furtado and Cunha, between 2013 and 2023 the number of data centres in the country expanded in response to demand for cloud services. This expansion did not generate technological autonomy, however, since the infrastructure installed remained under the control of international corporations operating according to a logic of accumulation and governance defined outside Brazil.[20]

According to a report produced by researchers at USP and UnB, between 2014 and 2025 the Brazilian state spent at least R$ 23 billion on software contracts, cloud services and security applications supplied by foreign corporations. Of this amount,

---

17 POLIDO, Fabrício Bertini Pasquot. Estado, soberania digital e tecnologias emergentes: interações entre direito internacional, segurança cibernética e inteligência artificial. *Revista de Ciências do Estado*, v. 9, n. 1, 2024.

18 MASSUCHIN; TAVARES; COMEL, cit.

19 FURTADO, Renato Guimarães; CUNHA, Simone Evangelista. Inteligência artificial, data centers e colonialismo digital: impactos socioambientais e geopolíticos a partir do Sul Global. *Liinc em Revista*, v. 20, n. 2, 2024. See also KWET, Michael. Digital colonialism: US empire and the new imperialism in the Global South. *Race & Class*, v. 60, n. 4, p. 3-26, 2019. DOI: 10.1177/0306396818823172.

20 FURTADO; CUNHA, cit.

R$ 10.35 billion was spent between June 2024 and June 2025 alone. Microsoft alone received more than R$ 3 billion, while Oracle, Google and Red Hat also figure among the main beneficiaries of public contracts. The dependence deepens year after year.[21]

These figures reveal more than a transfer of financial resources. They mark the consolidation of an architecture in which the Brazilian state functions as a mere consumer, subordinated to solutions controlled abroad. Such a configuration produces systemic vulnerabilities, since governmental and public data become, at least potentially, subject to legislation from outside the territory,[22] such as the U.S. CLOUD Act, examined in detail in the next section. The political economy of this dependence is not accidental. As Rikap has shown for Amazon, the leading cloud corporations operate as intellectual monopolies that convert technical and legal control over intangible assets into durable rent extraction from subordinated firms and states; once established, the dependence reproduces itself.[23]

The concept of digital colonialism is not limited to registering this technological dependence. It also addresses its geopolitical dimension: the continuous extraction of value from data produced in the Global South, processed and monetized by platforms located in the Global North. This is a deeply asymmetrical relation in which countries like Brazil supply informational raw material and consumer markets but are excluded from decisions concerning technological architectures and governance standards.[24]

A legal-institutional dimension completes this analysis: the absence of effective regulatory capacity over transnational platforms, itself the result of political choices that privilege corporate self-regulation over democratic control. Even advanced

---

21 SILVA, Ergon Cugler de Moraes; ROCHA, Isabela; VAZ, José Carlos; VENEZIANI, Julia Ribeiro de Almeida; MODANEZ, Camila de Camargo. *Contratos, códigos e controle: a influência das big techs no Estado brasileiro*. São Paulo: GETIP/EACH/USP; Brasília: GEPSI/IREL-UnB, jul. 2025. Disponível em: https://bit.ly/contratos-big-techs. Acesso em: 27 dez. 2025.

22 POLIDO, cit.

23 RIKAP, Cecilia. Amazon: a story of accumulation through intellectual rentiership and predation. *Competition & Change*, v. 26, n. 3-4, p. 436-466, 2022. DOI: 10.1177/1024529420932418.

24 COULDRY, Nick; MEJIAS, Ulises A. Data colonialism: rethinking big data's relation to the contemporary subject. *Television & New Media*, v. 20, n. 4, p. 336-349, 2019; FURTADO; CUNHA, cit.

legislation loses effectiveness against the economic power and lobbying of big tech.[25]

At this point, the structural limits of digital sovereignty in Global South countries call for a change of perspective: from the concept of sovereignty to that of hegemony, which illuminates dimensions of the phenomenon that would otherwise remain obscured. The Gramscian concept of hegemony refers to a form of power exercised through a combination of coercion and consensus, in which domination naturalizes itself to the point that explicit assertion becomes dispensable.[26]

U.S. technological hegemony does not manifest itself, at first sight, through overt impositions, but through the capacity to define the infrastructures, protocols and technological standards that shape the functioning of the global internet. This power operates less by interdiction than by formatting the conditions under which communication and the circulation of information take place.

Santaniello argues that it is where authority is weak that the discourse of sovereignty appears most intensely. This explains why the United States, one of the centres of global digital power, never robustly developed a discourse of digital sovereignty, notwithstanding salient exceptions such as the TikTok case. It is not necessary to assert sovereignty when one possesses hegemony. In this reading, it is the anti-sovereign, rather than the sovereign itself, that generates sovereignty as a discourse.[27] It is not U.S. hegemony that provokes claims of digital sovereignty, but the recognition that this hegemony is now contested.

From this perspective, the discourse of digital sovereignty is not a return to the Westphalian logic of territorially bounded states. It is a symptom of the rupture of a

---

[25] POLIDO, cit.

[26] COX, Gramsci, hegemony and international relations, cit.; GILL, Stephen; LAW, David. Global hegemony and the structural power of capital. *International Studies Quarterly*, v. 33, n. 4, p. 475-499, 1989. For its application to sovereignty claims in the digital sphere, SANTANIELLO, Mauro. *Attributes of digital sovereignty: a conceptual framework*. *Geopolitics*, v. 31, n. 2, p. 788-809, 2026 (publicado online em 2025). DOI: 10.1080/14650045.2025.2521548.

[27] SANTANIELLO, cit., p. 799. Santaniello's inversion extends a position now well established in the discursive-practice literature on digital sovereignty: see POHLE; THIEL, cit.; GLASZE, Georg et al. Contested spatialities of digital sovereignty. *Geopolitics*, v. 28, n. 2, p. 919-958, 2022; BROEDERS, Dennis; CRISTIANO, Fabio; KAMINSKA, Monica. In search of digital sovereignty and strategic autonomy. *JCMS: Journal of Common Market Studies*, v. 61, n. 5, p. 1261-1280, 2023.

previously established hegemony. Sovereignty discourses in the European Union, China and across the Global South indicate that the tacit consensus around the digital architecture designed by the United States is fragmenting, opening space for disputes.

This fragmentation is no longer bipolar. Bradford describes three regulatory empires competing to project their models globally: the U.S. market-driven, the European rights-driven and the Chinese state-driven model. A country like Brazil does not choose between sovereignty and dependence in the abstract; it navigates among three hegemonic projects.[28] China is the clearest example of the dispute. Chinese data policy has moved from a defensive posture focused on domestic sovereignty to an offensive strategy directed at constructing a hegemony of its own. Initiatives exporting technological infrastructure seek to establish alternatives that challenge the U.S. model.[29] Erie and Streinz describe this projection as a *"Beijing Effect"*: through the Digital Silk Road, China exports not only hardware and connectivity but also standards, contractual practices and governance models to recipient states, shaping transnational data governance from the infrastructure up.[30] What is at stake is the power to define the rules that will shape digital communication in the decades to come, not merely a dispute over markets.

Nor is the question simply one of physical infrastructure or technological capacity. Technological hegemony refers to the power to determine which information reaches which territory and population, which content obtains visibility, and which technological standards are adopted: the power to format the digital environment within which everyone else must act.[31] In the Brazilian case, this configuration is particularly problematic. The country faces a double subordination: dependent, on

---

28 BRADFORD, Anu. *Digital empires: the global battle to regulate technology*. New York: Oxford University Press, 2023.

29 BORGOGNO, Oscar; SAVINI ZANGRANDI, Michele. Chinese data governance and trade policy: from cyber sovereignty to the quest for digital hegemony? *Journal of Contemporary China*, v. 33, n. 148, p. 578-602, 2024.

30 ERIE, Matthew S.; STREINZ, Thomas. The Beijing Effect: China's Digital Silk Road as transnational data governance. *New York University Journal of International Law and Politics*, v. 54, n. 1, p. 1-91, 2021.

31 KWET, cit.; McCARTHY, The meaning of materiality, cit.

the one hand, on U.S. infrastructures and platforms, while witnessing, on the other, China's growing presence in strategic sectors. This double dependence leaves scant room for manoeuvre, placing the country squarely within the space of dispute between the two powers.[32]

The displacement from sovereignty to hegemony does not mean abandoning the first concept in favour of the second. It means recognizing that digital sovereignty, as claimed, is more a regulative horizon than an effective reality. Digital sovereignty will not be obtained through isolated technical solutions or national regulatory frameworks alone, but through strategies that combine technological development, international cooperation and the contestation of prevailing hegemonic structures; in Gramscian terms, a "war of position" waged on the terrain of infrastructure, law and standards.[33]

Having established the theoretical foundations for understanding U.S. technological hegemony and the digital colonialism that characterizes Brazil's subordinate insertion into the global informational economy, we can now examine the principal legal instrument through which this hegemony is given effect. The CLOUD Act is the clearest normative expression of the claim to extraterritorial jurisdiction over data processed by U.S. companies, regardless of physical location. Its analysis reveals not only technical-legal incompatibilities with the Brazilian legal order, but also the architecture of power that sustains the U.S. capacity to access information produced and stored outside its territory.

2 THE CLOUD ACT, FISA SECTION 702 AND THE LEGAL ARCHITECTURE OF EXTRATERRITORIAL ACCESS

International relations scholarship supplies a precise mechanism for what this section describes. Farrell and Newman have shown that states occupying the central nodes of global economic networks can convert that position into coercive capacity: they *"can employ the 'panopticon effect' to gather strategically valuable information"*

---

[32] POLIDO, cit.

[33] On counter-hegemony and the "war of position" in the international context, COX, cit.; GILL; LAW, cit.

and *"the 'chokepoint effect' to deny network access to adversaries"*.[34] In plain terms, a state at the centre of a network can watch what flows through it (the panopticon effect) and can cut others off from it (the chokepoint effect). The CLOUD Act is a legal expression of the first capacity; the sanctions and trade measures examined in section 3 express the second. The cloud is such a network, and U.S. law sits at its centre. The legal architecture supporting U.S. technological hegemony over data processed and stored abroad was consolidated through the Clarifying Lawful Overseas Use of Data Act (CLOUD Act), enacted by the U.S. Congress in 2018. The CLOUD Act gives statutory form to a claim of extraterritorial jurisdiction that challenges the foundations of other states' digital sovereignty. The statute amends prior legislation to establish that service providers are obliged to disclose information within their *"possession, custody, or control"*, regardless of whether such information is located inside or outside the United States.[35]

The novelty of the CLOUD Act lies not in the objective of reaching data outside U.S. territory, something that already occurred in criminal jurisdiction, but in the way the Act removes any doubt about what was previously debatable.[36] The law was prompted by *Microsoft Corp. v. United States*, in which an impasse had formed: Microsoft refused to hand over a user's data stored on an Irish server, arguing that the prior legislation did not authorize warrants beyond U.S. territory. Before the Supreme Court the case was restyled *United States v. Microsoft Corp.* and declared moot once the new law passed. The result was an expansion of U.S. investigative power, removing any doubt about the intention to subordinate the physical location of data to U.S. jurisdiction.

The structure of the CLOUD Act is articulated with another instrument of state surveillance, the Foreign Intelligence Surveillance Act (FISA). The former establishes modes of access to data for criminal investigations through judicial orders; Section

---

[34] FARRELL, Henry; NEWMAN, Abraham L. Weaponized interdependence: how global economic networks shape state coercion. *International Security*, v. 44, n. 1, p. 42-79, 2019. DOI: 10.1162/isec_a_00351, p. 45.

[35] UNITED STATES. CLOUD Act, cit. The quoted language is codified at 18 U.S.C. § 2713.

[36] COCHRANE, Tim. Hiding in the eye of the storm cloud: how CLOUD Act agreements expand U.S. extraterritorial investigatory powers. *Duke Journal of Comparative & International Law*, v. 32, n. 1, p. 47-92, 2021.

702 of FISA, in turn, authorizes the large-scale collection of electronic communications of persons located outside the United States for intelligence purposes. Although the CLOUD Act and FISA 702 have formally distinct purposes (law enforcement and intelligence, respectively), both permit the U.S. government to access data of foreigners stored by cloud providers.[37]

Section 702 deserves particular attention in the current conjuncture. Reauthorized and expanded in April 2024 by the Reforming Intelligence and Securing America Act (RISAA), the programme now reaches a broader class of *"electronic communication service providers"*, a category whose precise contours remain partly classified. [38] The Privacy and Civil Liberties Oversight Board (PCLOB), the independent body charged with overseeing such programmes, documented in 2014 the structural features of Section 702 collection.[39] The removal of three of its members in January 2025, however, left the Board without a quorum, so its most recent assessment, in April 2026, could be issued only as a staff report, without the authority of the Board itself. Independent oversight of U.S. surveillance programmes is, for now, weakened. [40] For a non-U.S. person, whether a Brazilian citizen, court or ministry, Section 702 confers no judicially enforceable rights: it is designed to collect the communications of foreigners. RISAA did introduce internal reforms, notably restrictions and approval requirements for FBI queries of U.S.-person information and reinforced

---

37 MICHELS, Johan David; MILLARD, Christopher; WALDEN, Ian; WUERMELING, Ulrich. *Cloud sovereignty and the GDPR, Part One: US government access to European data*. Queen Mary Law Research Paper No. 431/2024. London: Queen Mary University of London, 2024. Disponível em: https://ssrn.com/abstract=4911552. Acesso em: 27 dez. 2025.

38 CONGRESSIONAL RESEARCH SERVICE (CRS). *FISA Section 702 and the 2024 RISAA*. CRS Report R48592. Washington, DC: Library of Congress, 2025. Disponível em: https://crsreports.congress.gov/product/pdf/R/R48592. Acesso em: 31 jul. 2026.

39 PRIVACY AND CIVIL LIBERTIES OVERSIGHT BOARD (PCLOB). *Report on the surveillance program operated pursuant to Section 702 of the Foreign Intelligence Surveillance Act*. Washington, DC: PCLOB, 2 jul. 2014.

40 PRIVACY AND CIVIL LIBERTIES OVERSIGHT BOARD (PCLOB). *Report on the surveillance program operated pursuant to Section 702 of the Foreign Intelligence Surveillance Act: staff report*. Washington, DC: PCLOB, 2 abr. 2026; BRENNAN CENTER FOR JUSTICE. *Section 702 of FISA explained*. New York: Brennan Center for Justice, 2026. Disponível em: https://www.brennancenter.org/our-work/research-reports/section-702-fisa-explained. Acesso em: 31 jul. 2026.

minimization procedures, but these discipline domestic querying of the collected data. They do not create procedural protection for the foreign persons whose communications are collected.[41]

Michels, Millard, Walden and Wuermeling identify that U.S. jurisdiction attaches not to the territory where data is physically stored, but to the legal person that has control over it, whether direct technical control (de facto control) or legal control through corporate structure (legal control). A company headquartered in the United States, such as Amazon, Microsoft or Google, remains subject to U.S. legislation even when operating data centres in Europe, Brazil or any other jurisdiction.[42]

The traditional objective criterion of physical location is thereby replaced by a subjective criterion: corporate nationality or the seat of the controller. The consequence is that it does not matter whether the data of a Brazilian citizen, processed by AWS, is physically stored on servers in São Paulo. A request addressed to AWS in Delaware produces its effects in full.

This logic articulates with what Zuboff terms surveillance capitalism. Companies such as Google, Meta and Amazon have developed a new logic of accumulation based on the exploitation of behavioural data. The CLOUD Act can be read as the state instrumentalization of this commercial surveillance infrastructure: while big tech extracts economic value from data, the state extracts geopolitical value from the same base, since the massive concentration of information about foreign populations is accumulated by companies subject to the U.S. government and accessible to it.[43] A state–capital symbiosis is thus renewed, in which the economic dominance of big tech sustains and is sustained by U.S. dominance over global digital infrastructure. The relationship is not one of simple instrumentality, however. The same corporations have at times litigated against U.S. secrecy orders and, in 2025, publicly committed to challenging government demands threatening their European operations. Big tech thus has interests of its own, though commitments of this kind do not release a company from complying with a lawful order.[44]

---

41 CRS, R48592, cit.

42 MICHELS et al., Cloud sovereignty and the GDPR, cit., p. 17-18.

43 ZUBOFF, cit.

44 MICROSOFT. European digital commitments. Redmond, 30 Apr. 2025. Disponível em:

Analyzing the compatibility of this regime with the European General Data Protection Regulation (GDPR), Michels and colleagues examine four potential compliance problems. The problems were formulated for the GDPR; we transpose them, with the necessary adaptations, to the Brazilian General Data Protection Law (LGPD). First, U.S. government access may constitute processing without the controller's instructions, since providers hand personal data to foreign authorities without any instruction from the data controller. Second, there may be no adequate legal basis, since the GDPR requires processing to be grounded in EU or Member State law, and does not recognize third-country legislation as a legal basis. Third, there is a risk of violation of purpose limitation, as data lawfully collected for commercial purposes is diverted to criminal investigation or intelligence unconnected with the original purpose. Fourth, security obligations may be breached, since the cloud provider, in complying with CLOUD Act orders, enables access to personal data by third parties without the original controller's authorization.[45]

The LGPD and the GDPR share a similar structure for international data transfers: both rest on the principle of adequacy, under which personal data may only be transferred to third countries offering a level of protection equivalent to that of the originating legislation. International transfers, in other words, must not weaken the protection afforded to data in the territory of origin.

There is, however, a crucial difference between the two laws. The GDPR contains, in its Article 48, a provision that part of the international scholarship characterizes as functioning like a blocking statute: a norm under which foreign judgments or administrative decisions ordering the transfer of personal data are not, in themselves, recognized as a lawful ground for transfer, save where grounded in an international treaty or mutual legal assistance agreement. According to Swire, in certain circumstances Article 48 *"clearly acts as a blocking statute"*, especially where no prior legal basis for the transfer exists.[46] The European Data Protection Board consolidated this

https://blogs.microsoft.com/on-the-issues/2025/04/30/european-digital-commitments/. Acesso em: 1 ago. 2026.

45 MICHELS et al., Cloud sovereignty and the GDPR, cit.

46 SWIRE, Peter. *When does GDPR act as a blocking statute: the relevance of a lawful basis for transfer*. Georgia Tech Scheller College of Business Research Paper, 2019. Disponível em: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3473187. Acesso em: 31 dez. 2025.

reading in its Guidelines 02/2024 on Article 48, adopted in final version on 4 June 2025 (with a technical update, Version 2.1, on 20 June 2025). A foreign authority's order is not, in itself, a lawful ground for transfer, although the EDPB is equally clear that Article 48 is a rule on the recognition of foreign decisions rather than a complete blocking statute in the classic sense.[47] The LGPD contains no equivalent provision, which leaves it uncertain whether companies operating in Brazil must comply with foreign orders.

A procedural incompatibility compounds the material one. The CLOUD Act partly codified the previously judge-made comity doctrine: under 18 U.S.C. § 2703(h), a provider may move to quash or modify qualifying U.S. process where the customer is not a U.S. person and disclosure would create a material risk of violating the law of a *"qualifying foreign government"*, that is, a government with which the United States has concluded an executive agreement under § 2523.[48] For countries without such an agreement, Brazil among them, only the traditional common-law comity analysis remains available. As Cochrane emphasizes, that analysis offers fragile protection in practice. The burden falls on the provider to demonstrate that a legal conflict exists, that the foreign law applies, and that the foreign country's interests outweigh those of the United States, before U.S. judges applying U.S. factors; few comity challenges succeed.[49]

The result is an asymmetry. The GDPR affords stronger protection, and qualifying-government providers enjoy a statutory motion; the absence of any such provision in the LGPD, combined with Brazil's position outside the § 2523 club, leaves Brazilian companies and especially public bodies using cloud services from U.S.-based providers exposed to extraterritorial jurisdiction.

The reality of this risk was made explicit in the public hearing conducted in the French Senate with executives of Microsoft France, before the committee of inquiry

---

47 EUROPEAN DATA PROTECTION BOARD (EDPB). *Guidelines 02/2024 on Article 48 GDPR*. Version 2.0, adopted 4 June 2025; Version 2.1, 20 June 2025. Brussels: EDPB, 2025. Disponível em: https://www.edpb.europa.eu/system/files/2025-06/edpb_guidelines_202402_article48_v2_en.pdf. Acesso em: 31 jul. 2026.

48 UNITED STATES. 18 U.S.C. § 2703(h). Disponível em: https://www.law.cornell.edu/uscode/text/18/2703. Acesso em: 1 ago. 2026.

49 COCHRANE, cit.

into public procurement and its knock-on effects on the French economy, hearings in which digital sovereignty took centre stage. The testimony given on 10 June 2025 by Anton Carniaux, Director of Public and Legal Affairs of Microsoft France, accompanied by Pierre Lagarde, Technical Director for the Public Sector, gave concrete form to what had remained a theoretical concern of jurists: companies headquartered in the United States cannot guarantee the data sovereignty of other countries in the face of the CLOUD Act.[50]

Asked by the committee's rapporteur whether he could guarantee under oath that data of French citizens stored in Microsoft's European data centres would never be transferred to U.S. authorities without the approval of French authorities, Carniaux answered: *"Non, je ne peux pas le garantir, mais, encore une fois, cela ne s'est encore jamais produit" — "No, I cannot guarantee that, but, once again, that has never yet happened"*.[51] What the testimony establishes is the express corporate acknowledgment that, were such an order to come, contractual and technical safeguards would not eliminate the legal obligation to comply; to Microsoft's knowledge, no such disclosure of European public-sector data has yet occurred. Lagarde added that, on the technical side, the services at issue remain within the European Union, and that data is encrypted at rest and in transit under Microsoft's contractual commitments; press coverage of the hearing drew the same conclusion on the legal exposure.[52]

For Brazil, this testimony matters because it shows that the exposure created by the CLOUD Act is real and acknowledged by the big tech companies themselves. One caveat is owed to the reader. No public record located in the research for this study

---

[50] FRANCE. Sénat. Commission d'enquête sur les coûts et les modalités effectifs de la commande publique et la mesure de son effet d'entraînement sur l'économie française. Audition de MM. Anton Carniaux, directeur des affaires publiques et juridiques, et Pierre Lagarde, directeur technique du secteur public, Microsoft France (présidence de M. Simon Uzenat). Séance du mardi 10 juin 2025. Compte rendu disponível em: https://www.senat.fr/compte-rendu-commissions/20250609/ce_commande_publique.html. Acesso em: 1 ago. 2026.

[51] FRANCE. Sénat, Audition de représentants de Microsoft France, cit.

[52] PPC LAND. Microsoft can't protect French data from US government access. 18 July 2025. Disponível em: https://ppc.land/microsoft-cant-protect-french-data-from-us-government-access/. Acesso em: 26 dez. 2025; THE REGISTER. Microsoft exec admits it 'cannot guarantee' data sovereignty. 25 July 2025. Disponível em: https://www.theregister.com/off-prem/2025/07/25/microsoft_admits_it_cannot_guarantee/. Acesso em: 31 jul. 2026.

documents a CLOUD Act order actually served on a provider for Brazilian public-sector data and complied with. Orders and provider responses may be sealed or aggregated in transparency reporting, so the absence of a public case does not establish the absence of access; still, the distinction between demonstrated statutory exposure and documented realized harm should be kept explicit.[53] The flagship Brazilian cases examined next are, accordingly, cases about procurement, institutional competence and cyber-governance risk under conditions of statutory exposure, and they are no less instructive for that. Having explored the abstract mechanisms of extraterritorial jurisdiction, we now turn to how this architecture manifests itself in the Brazilian context.

3 BRAZILIAN DIGITAL DEPENDENCE: MANIFESTATIONS, RISKS AND INSTITUTIONAL RESPONSES

Having explored the legal architecture that sustains U.S. technological hegemony, we now investigate how it manifests itself in the Brazilian context, for if the effects are material, the necessary responses cannot be abstract. Brazil's dependence on infrastructure controlled by U.S. and other foreign companies is a real structural vulnerability before mechanisms such as the CLOUD Act. Three episodes organize this section: the São Paulo Court of Justice contract of 2019, a procurement interrupted on sovereignty grounds; the DATASUS migration of 2020 and the cyberattack of 2021, in which sensitive data was moved under conditions of exposure; and the sanctions and trade offensive of 2025–2026, in which network centrality was exercised openly against Brazilian institutions. The first two evidence latent risk. The third shows the same architecture in active operation.

---

[53] UNITED STATES. Department of Justice. CLOUD Act resources. Washington, DC. Disponível em: https://www.justice.gov/criminal/cloud-act-resources. Acesso em: 1 ago. 2026; CONGRESSIONAL RESEARCH SERVICE (CRS). Law enforcement access to overseas data under the CLOUD Act. LSB10125. Washington, DC: Library of Congress. Disponível em: https://www.congress.gov/crs-product/LSB10125. Acesso em: 1 ago. 2026.

3.1 The São Paulo Court of Justice and Microsoft

The first case occurred in February 2019, when the São Paulo Court of Justice (TJ-SP), the state court of São Paulo and the largest judicial system in Brazil, announced a contract of R$ 1.32 billion with Microsoft for the implementation of the so-called Plataforma Lex, migrating its case-management system to Microsoft Azure infrastructure and storing judicial data on remote servers. The procurement was conducted with waiver of public bidding under the Technological Innovation Law (Law No. 10,973/2004, which permits non-competitive contracting for research and innovation projects), on the argument that the partnership constituted a research, development and innovation project.[54]

The National Council of Justice (CNJ), the constitutional body charged with administrative and financial oversight of the Brazilian judiciary (Constitution, Article 103-B, § 4), reacted only one day after the announcement, suspending the contract by preliminary decision and pointing to procedural irregularities and risks to national sovereignty.[55] Three main grounds motivated the decision. First, the legality of the bidding waiver: although the Technological Innovation Law permits direct contracting for innovation projects, the mere migration of judiciary systems to a commercial cloud would not fit this hypothesis. Second, a conflict with CNJ Resolution No. 185/2013, which establishes the Electronic Judicial Process (PJe) as the case-management system to be prioritized in Brazil, incompatible with the adoption of an alternative platform. Third, and most relevant to our argument, the decision invoked risks to national security in allowing a *"foreign company, on foreign soil, [to] keep custody of and access to judicial data of Brazil"*, which would represent an unacceptable vulnerability.[56]

The national sovereignty argument gained force when a potential violation of the

---

54 CONSULTOR JURÍDICO. CNJ suspende contrato de R$ 1,3 bi entre TJ-SP e Microsoft. 21 fev. 2019. Disponível em: https://www.conjur.com.br/2019-fev-21/cnj-suspende-contrato-13-bi-tj-sp-microsoft. Acesso em: 27 dez. 2025.

55 BRASIL. Constituição da República Federativa do Brasil de 1988, art. 103-B, § 4º. Brasília: Presidência da República, 1988.

56 CONSULTOR JURÍDICO, CNJ suspende contrato, cit.

LGPD, not yet fully in force at the time, was raised, since its Article 26 restricts the transfer of data by public bodies to private entities, with limited exceptions. On 9 April 2019 the CNJ plenary voted 14–1 (with a single dissent by Councillor Luciano Frota) to maintain the suspension while permitting technical studies, and on 25 June 2019 it unanimously denied the TJ-SP's appeal. The following year the TJ-SP rescinded the contract with Microsoft, maintaining only a limited partnership for the use of Microsoft Teams.[57]

Two precisions frame what the episode does and does not establish. It was an exercise of the CNJ's administrative and financial oversight of a state court's procurement, not an adjudication that a CLOUD Act order had violated, or would violate, Brazilian sovereignty. For that very reason it supports our argument: a domestic governance mechanism proved capable of interrupting a dependency trajectory on sovereignty grounds, while leaving the foreign-order question open. The lessons remain: contractual guarantees alone do not protect data sovereignty when the provider is subject to a foreign state; dependence on foreign infrastructure for critical state services raises legitimate legal concerns; and the absence of viable alternatives places the Brazilian state at a crossroads between modernization and the preservation of autonomy.

The legal analysis shows how the four compliance problems examined by Michels and colleagues would manifest concretely. First, processing without the controller's instructions: the TJ-SP, controller under Article 5, VI, of the LGPD, would lose the capacity to determine the purposes and means of processing if the processor (Article 5, VII) were compelled by U.S. authorities to provide access. The transfer of data to third parties without the controller's express instruction directly violates Article 39 of the LGPD, under which processors may only process data according to the controller's instructions. Second, the absence of an adequate legal basis: Article 7 of the LGPD, which enumerates the hypotheses of lawful processing, does not contemplate

[57] MIGALHAS. CNJ mantém suspensa contratação do TJ-SP com Microsoft, mas permite estudos técnicos. 9 abr. 2019. Disponível em: https://www.migalhas.com.br/quentes/300031/cnj-mantem-suspensa-contratacao-do-tj-sp-com-microsoft--mas-permite-estudos-tecnicos. Acesso em: 1 ago. 2026; CONSULTOR JURÍDICO. TJ-SP rescinde contrato de R$ 1,3 bilhão com a Microsoft. 20 maio 2020. Disponível em: https://www.conjur.com.br/2020-mai-20/tj-sp-rescinde-contrato-13-bilhao-microsoft/. Acesso em: 27 dez. 2025.

compliance with a foreign judicial order. Third, violation of purpose limitation under Article 6, I, of the LGPD: judicial data is collected and processed for specific purposes, and its sharing with foreign authorities constitutes a deviation of purpose incompatible with the original legitimacy of the processing; the purpose cannot be altered without a new legal basis, which would not exist in a CLOUD Act request. Fourth, security obligations: Article 46 of the LGPD requires processors to implement technical and administrative measures to protect data against unauthorized access.[58]

3.2 DATASUS and the health data of the Brazilian population

The second case concerns the Department of Informatics of the Unified Health System (DATASUS), the information-technology arm of Brazil's national public health system, responsible for processing data on the entire Brazilian population. It shows the vulnerability created by transferring sensitive public health data to infrastructure controlled by a U.S. company, reaching the most sensitive category protected by the LGPD. In April 2020, according to trade-press reports at the time, the Ministry of Health concluded a contract with Embratel to migrate DATASUS data to the Amazon Web Services (AWS) cloud: millions of medical records of Brazilians, including care histories, hospitalizations, medications and vaccination data, moved to servers operated by a company subject to the CLOUD Act. The same reporting describes the project as part of a broader procurement of cloud computing services involving some 23 federal bodies, including CADE (the competition authority), the CNJ and the INSS (the social security institute); no primary government record of these figures is public, so we attribute them to that reporting.[59]

This migration was itself embedded in a wider shift in Brazilian digital government policy. In 2016, the Temer administration created the Digital Citizenship Platform through Decree No. 8,936; from 2019–2020 the federal government formal-

---

58 MICHELS et al., Cloud sovereignty and the GDPR, cit.; BRASIL. Lei nº 13.709, de 14 de agosto de 2018 (Lei Geral de Proteção de Dados Pessoais — LGPD). Brasília: Presidência da República, 2018.

59 BAGUETE. Ministério da Saúde migra para AWS com Embratel. 22 abr. 2020. Disponível em: https://www.baguete.com.br/noticias/22/04/2020/ministerio-da-saude-migra-para-aws-com-embratel. Acesso em: 31 dez. 2025.

ized a *"cloud first"* orientation for the federal administration, an administrative directive of the Digital Government Secretariat under Decree No. 10,332/2020, instructing SISP bodies to prioritize cloud solutions over local infrastructure where secure, reliable and economical. The directive did not itself mandate contracting multinational providers. In practice, however, given the market structure described in section 1, prioritizing commercial cloud meant prioritizing foreign hyperscalers.[60]

What was presented as technological modernization carried serious implications for digital sovereignty and the protection of fundamental rights. Health data is a special category protected by the LGPD, particularly under Article 11, demanding heightened care by reason of its potential discriminatory and improper use. The law establishes that such processing may only occur with the specific consent of the data subject or legal guardian, or in clearly defined legal hypotheses, always with security guarantees. By transferring DATASUS data to AWS, the Brazilian state placed legally protected information under the potential jurisdiction of foreign authorities via the CLOUD Act or FISA, as discussed above.

Public health information may be of interest both for individual criminal investigations and for its strategic value to state and corporate intelligence. As the PRISM programme revealed in 2013, U.S. intelligence agencies have a documented history of massive collection of foreign citizens' data through technology platforms.[61] DATASUS data, containing the epidemiological mapping of the population, disease registries and patterns of medication use, is an obvious target for health and economic intelligence.

The vulnerability materialized in December 2021, when the Ministry of Health

---

[60] BRASIL. Secretaria de Governo Digital (SGD). O que é a diretriz Cloud First da SGD para o SISP? Brasília. Disponível em: https://www.gov.br/governodigital/pt-br/estrategias-e-governanca-digital/estrategias-e-politicas-digitais/computacao-em-nuvem/o-que-e-a-diretriz-cloud-first-da-sgd-para-o-sisp. Acesso em: 1 ago. 2026; COLI, Anna Luiza; AVRITZER, Gabriel Caetano. Da promessa da soberania digital à realidade do aluguel tecnológico. Belo Horizonte: UFMG, 5 set. 2025. Disponível em: https://www.ufmg.br/comunicacao/noticias/opiniao/da-promessa-da-soberania-digital-a-realidade-do-aluguel-tecnologico/. Acesso em: 31 dez. 2025.

[61] GREENWALD, Glenn; MACASKILL, Ewen. NSA Prism program taps in to user data of Apple, Google and others. *The Guardian*, London, 6 jun. 2013. Disponível em: https://www.theguardian.com/world/2013/jun/06/us-tech-giants-nsa-data. Acesso em: 3 ago. 2026.

suffered a cyberattack that struck the Ministry's website, ConecteSUS, e-SUS Notifica and the National Immunization Programme Information System, in the middle of the Covid-19 pandemic.[62] The attack hit private cloud infrastructure contracted in 2018 from Primesys, an Embratel subsidiary, by the Ministry of Planning. Subsequent investigation revealed serious contractual fragilities: according to the company itself, the agreement with the Brazilian government did not include security services, only hosting.[63]

The four categories of violation can again be applied to the DATASUS case. Processing without the controller's instructions manifests itself where AWS, operating as processor (Article 5, VII, LGPD) under the instructions of the Ministry of Health (controller), is compelled by a U.S. order to provide health data without the Brazilian controller's authorization or knowledge. The absence of an adequate legal basis is especially problematic here: Article 11, § 4, of the LGPD expressly restricts the communication or shared use of sensitive health data between controllers for purposes other than those for which it was collected. The violation of purpose limitation is evident, since epidemiological information, vaccination records and medical histories were collected for public health purposes (Article 7, II and III, together with Article 11, II, *"f"*), not to feed the criminal or intelligence investigations of a foreign government. And the security obligations are critical for health data: Article 46, § 1, of the LGPD requires measures apt to protect data against accidental or unlawful destruction, loss, alteration, communication or unauthorized access. From the Brazilian perspective, a U.S. request constitutes unauthorized access, however valid it may be under U.S. law.

---

62 AGÊNCIA BRASIL. Sites e aplicativo do Ministério da Saúde sofrem ataque cibernético. 10 dez. 2021. Disponível em: https://agenciabrasil.ebc.com.br/saude/noticia/2021-12/sites-e-aplicativo-do-ministerio-da-saude-sofrem-ataque-cibernetico. Acesso em: 27 dez. 2025.

63 BRASIL DE FATO. ConecteSUS fora do ar: governo contratou nuvem privada sem serviço de segurança. São Paulo, 21 dez. 2021. Disponível em: https://www.brasildefato.com.br/2021/12/21/conectesus-fora-do-ar-governo-contratou-nuvem-privada-sem-servico-de-seguranca. Acesso em: 31 dez. 2025.

3.3 From latent risk to open confrontation: the 2025–2026 sanctions and trade offensive

If the TJ-SP and DATASUS cases evidence latent risk, the events of 2025–2026 show what Farrell and Newman's framework predicts: an actor that occupies the chokepoints of financial, technological and trade networks can convert that centrality into targeted coercion.[64] Three distinct legal tracks, often blurred in public debate, must be kept separate.

The first track was financial. On 30 July 2025, the U.S. Treasury designated a sitting Justice of the Supreme Federal Tribunal, Alexandre de Moraes, under the Global Magnitsky programme, in explicit connection with judicial proceedings against U.S.-based platforms and the criminal trial of a former president; his wife was added to the sanctions list in September 2025, and both designations were revoked on 12 December 2025.[65] While they lasted, the sanctions produced immediate compliance dilemmas for banks and payment processors operating in Brazil, squeezed between U.S. sanctions law and the orders of Brazilian courts. As Brazilian commentary observed, the episode projected the U.S. legal order directly into Brazilian territory through the private infrastructure, financial and technological, on which Brazilian institutions depend.[66]

The second track was tariff-based and rested on emergency powers. By executive order of 30 July 2025, invoking the International Emergency Economic Powers Act and the National Emergencies Act, the United States imposed an additional 40%

---

64 FARRELL; NEWMAN, cit. See also, for the extension of the argument to cloud computing and semiconductors, FARRELL, Henry; NEWMAN, Abraham L. *Underground empire: how America weaponized the world economy*. New York: Henry Holt, 2023 (a trade monograph, cited here as illustrative).

65 MIGALHAS. EUA revogam sanções da lei Magnitsky contra Moraes e esposa. dez. 2025. Disponível em: https://www.migalhas.com.br/quentes/446354/eua-revogam-sancoes-da-lei-magnitsky-contra-moraes-e-esposa. Acesso em: 31 jul. 2026; AGÊNCIA BRASIL. "Verdade prevaleceu", diz Moraes sobre fim de sanções. Brasília, dez. 2025. Disponível em: https://agenciabrasil.ebc.com.br/justica/noticia/2025-12/verdade-prevaleceu-diz-moraes-sobre-fim-de-sancoes. Acesso em: 31 jul. 2026.

66 CONSULTOR JURÍDICO, Lei Magnitsky projeta ordem jurídica dos EUA, cit.

tariff on Brazilian goods, bringing the total, with the pre-existing 10% reciprocal tariff, to 50%. The measure was publicly linked to the conduct of Brazilian courts and regulators towards U.S. technology companies and to the trial of the former president.[67]

The third track was the formal trade instrument. On 15 July 2025 the U.S. Trade Representative directed a Section 301 investigation into Brazil's *"acts, policies, and practices"* in six areas: digital trade and electronic payment services (including judicial decisions applied to U.S. platforms and the alleged disadvantaging of U.S. companies by Pix, the Central Bank's instant payment system), preferential tariffs, anti-corruption enforcement, intellectual property, ethanol market access and deforestation.[68] On 1 June 2026, the USTR determined that certain of these practices were *"unreasonable"* and actionable, and in July 2026 announced a 25% tariff action under the Section 301 framework.[69]

These episodes should be read with care. They do not prove a single coordinated U.S. strategy uniting Treasury, the White House and USTR around digital policy; each track has its own statutory logic, constituency and reversibility, as the December 2025 revocation of the sanctions itself shows. What they demonstrate, cumulatively, is the observed effect that matters here. The exercise of Brazilian jurisdiction over digital platforms, and the construction of Brazilian public digital infrastructure, figured among the conducts drawing coercive responses through the network chokepoints the United States controls. Pix is itself an instance of sovereign public infrastructure, and the investigation treats it as a trade irritant in part because it competes

---

67 UNITED STATES. Executive Order of 30 July 2025 (Brazil tariffs). *Federal Register*, Washington, DC, 5 Aug. 2025. Disponível em: https://www.federalregister.gov/documents/full_text/xml/2025/08/05/2025-14896.xml. Acesso em: 1 ago. 2026.

68 UNITED STATES. USTR. Initiation of Section 301 Investigation, cit.

69 UNITED STATES. USTR. USTR Section 301 Determination on Brazil's Unreasonable Acts, Policies, and Practices. Washington, DC, 1 June 2026. Disponível em: https://ustr.gov/about/policy-offices/press-office/press-releases/2026/june/ustr-section-301-determination-brazils-unreasonable-acts-policies-and-practices. Acesso em: 31 jul. 2026; UNITED STATES. USTR. USTR Section 301 Action on Brazil's Unreasonable Acts, Policies, and Practices. Washington, DC, 14 July 2026. Disponível em: https://ustr.gov/about/policy-offices/press-office/press-releases/2026/july/ustr-section-301-action-brazils-unreasonable-acts-policies-and-practices. Acesso em: 1 ago. 2026.

with U.S.-controlled payment networks.[70] The cost of dependence is no longer hypothetical. It is the lever through which external pressure is applied.

3.4 Institutional responses: the sovereign cloud, the AI bill and the data centre policy

The vulnerability of DATASUS is a symptom of a broader structural dependence. As noted, the Brazilian public sector spent at least R$ 23 billion between 2014 and 2025 on technology services from foreign corporations, with Microsoft holding the largest contract value. The amount paid would suffice to build and inaugurate 86 Tier 3 data centres of 5 MW on national territory, increasing the country's current capacity by 50%.[71]

Aware of the gravity of this dependence, the federal government began investing from 2023 onwards in the construction of the so-called *"sovereign cloud"* (Nuvem de Governo), operated by SERPRO, the federal government's data-processing company, and Dataprev, with the objective of repatriating public administration information that had been transferred to foreign cloud providers.[72] The investment envelope is of approximately R$ 1 billion, of which R$ 710 million is executed through SERPRO under the Brazilian AI Plan.[73] By mid-2026, SERPRO reported hosting more than five thousand government systems in the sovereign cloud, presenting the initiative as a consolidated state policy for digital sovereignty.[74]

---

70 MOURA, Stéfano Mariotto de. "Economia política dos dados" como alternativa à "governança da internet": uma perspectiva desde o sul global. *Brazilian Journal of International Relations*, v. 14, e025021, 2025. DOI: 10.36311/2237-7743.2025.v14.e025021.

71 SILVA et al., cit.

72 BRASIL. Ministério da Gestão e da Inovação em Serviços Públicos (MGI). Gestão assina acordos com Serpro e Dataprev para criação de Nuvem de Governo. Brasília, 12 nov. 2024; BRASIL. MGI. Gestão, Serpro e Dataprev lançam serviços de nuvem de governo para órgãos federais. Brasília, 10 jun. 2025. Disponível em: https://www.gov.br/governodigital/pt-br/noticias/gestao-serpro-e-dataprev-lancam-servicos-de-nuvem-de-governo-para-orgaos-federais. Acesso em: 31 jul. 2026.

73 SERPRO. Nuvem Serpro e inteligência artificial nacional. Brasília, 31 jul. 2024. Disponível em: https://www.serpro.gov.br/menu/noticias/noticias-2024/nuvem-serpro-inteligencia-artificial-nacional. Acesso em: 1 ago. 2026.

74 MOBILE TIME. Serpro: nuvem soberana hospeda 5 mil sistemas do governo e 50 terabytes. 11

The sovereign cloud keeps information on Brazilian territory. The initiative is a significant advance, but it does not fully resolve the vulnerability arising from legislation such as the CLOUD Act. Migration is gradual, the country remains deeply dependent on proprietary software licensed from foreign companies, and the infrastructure itself incorporates foreign technology. According to SERPRO's own account, when it sought a cloud solution with fully isolated infrastructure, without external orchestration, the only supplier initially able to deliver it was China's Huawei, whose partnership with SERPRO was confirmed in response to a parliamentary information request.[75]

This fact must be confronted rather than glossed over. Chinese law imposes on Chinese organizations duties of support, assistance and cooperation with state intelligence work (National Intelligence Law, Article 7), applies to certain data-processing activities outside China, requires cooperation with domestic security demands, and prohibits supplying China-stored data to foreign authorities without Chinese approval (Data Security Law, Articles 2, 35 and 36).[76] This is not a mirror image of the CLOUD Act: the jurisdictional predicates, institutions, transparency and remedies differ, and no evidence suggests actual Chinese access to Brazilian government data. It is, however, a distinct and less transparent exposure. The lesson is architectural. Substituting a non-U.S. supplier does not, by itself, produce sovereignty. What matters is who controls the cryptographic keys, the orchestration layer, privileged access, auditing, update pathways and exit options. Diversification that ignores these

maio 2026. Disponível em: https://www.mobiletime.com.br/noticias/11/05/2026/serpro-nuvem-soberana/. Acesso em: 31 jul. 2026; AGÊNCIA GOV. Soberania digital: Brasil consolida a Nuvem de Governo. Brasília, set. 2025. Disponível em: https://agenciagov.ebc.com.br/noticias/202509/na-soberania-digital-brasil-consolida-a-nuvem-de-governo-como-passo-para-a-independencia-e-a-protecao-de-dados. Acesso em: 31 jul. 2026.

75 MOBILE TIME, cit.; BRASIL. Câmara dos Deputados. Resposta do SERPRO ao Requerimento de Informação nº 108/2025. Brasília, 2025. Disponível em: https://www.camara.leg.br/proposicoesWeb/prop_mostrarintegra?codteor=2888474&filename=Tramitacao-18-RIC-108-2025. Acesso em: 1 ago. 2026.

76 CHINA. National Intelligence Law of the People's Republic of China (2017, as amended), art. 7; CHINA. Data Security Law of the People's Republic of China (2021), arts. 2, 35-36. Disponível em: http://www.npc.gov.cn/englishnpc/c2759/c23934/202112/t20211209_385109.html. Acesso em: 1 ago. 2026.

dimensions may reduce one chokepoint while creating another, reproducing inside the sovereign cloud the double subordination described in section 1. Meanwhile, while more than R$ 1 billion was invested in the sovereign cloud, the public sector spent R$ 10.35 billion on foreign technology between June 2024 and June 2025 alone.[77]

The technical dimension of sovereignty deserves emphasis, because it partially decouples protection from localization. With client-side encryption and externally held or customer-held keys, hardware security modules, and confidential-computing techniques that protect data in use, a provider can be placed in a position where it cannot produce intelligible data even under compulsion; joint NSA and CISA guidance itself notes that with client-side encryption and customer-managed keys the provider lacks the keys needed to decrypt the data it stores.[78] In other words, if only the government customer holds the key, the provider has nothing intelligible to hand over. These controls are not complete legal immunity, since metadata, backups, misconfiguration, compelled assistance and any provider-held key path remain exposed, but they suggest a control hierarchy for policy design: customer-held keys and confidential computing where feasible; localization on sovereign infrastructure, or disconnected operation, for the residual highest-risk workloads. Put simply: encryption the provider cannot break wherever that is enough, and state-run or disconnected infrastructure for the most sensitive systems. The four-axis agenda in the conclusion adopts this hierarchy.

The legislative front reveals a pattern of sovereignty invoked but not operationalized. Bill No. 2,338/2023, which seeks to regulate AI in Brazil, was approved by the Federal Senate in December 2024 and, since a special commission was constituted in the Chamber of Deputies in 2025, has awaited the rapporteur's report, with a floor

---

77 SILVA et al., cit.

78 UNITED STATES. National Security Agency (NSA); Cybersecurity and Infrastructure Security Agency (CISA). *Use secure cloud key management practices*: cybersecurity information sheet. Washington, DC: NSA; CISA, mar. 2024. Disponível em: https://media.defense.gov/2024/Mar/07/2003407858/-1/-1/0/CSI-CloudTop10-Key-Management.PDF. Acesso em: 3 ago. 2026; MICROSOFT. Controls and principles in sovereign public cloud. Redmond, 2026. Disponível em: https://learn.microsoft.com/en-us/azure/azure-sovereign-clouds/public/overview-controls-principles. Acesso em: 1 ago. 2026.

vote expected only at the end of 2026.[79] The bill adopts a risk-based approach inspired by the European AI Act, recognizing as high-risk the systems used in recruitment, credit scoring, education and health, among others. It leaves a relevant gap, however: it does not confront the question of data sovereignty. Although it establishes data governance obligations, the bill is silent as to where data must be stored and who controls the infrastructure that processes it. Leading Brazilian scholars have argued in the same direction: the AI framework should be crafted as a vector of sovereignty, neither a copy of the European model nor of deregulatory alternatives.[80]

The bill mentions technological sovereignty explicitly as an objective, but only in the abstract, without providing for its operationalization, such as a preference for national infrastructure or mechanisms of protection against extraterritorial access. This absence reflects political and economic tensions, including intense pressure from industry. Whatever its causes, the prolonged passage through the Chamber keeps the window for correction open: the special commission has been the stage of a dispute over how much sovereignty the Brazilian AI framework will contain.

The same tension traverses the two other pillars of current Brazilian AI policy. The Brazilian Artificial Intelligence Plan (PBIA 2024–2028) promises investments of R$ 23.03 billion and elects sovereign AI infrastructure as a strategic axis, including national compute capacity.[81] And the REDATA bill (PL No. 278/2026), approved by the Chamber of Deputies and before the Senate since 25 February 2026, creates a special tax regime to attract data centre investment, conditioned on renewable energy use. Its trajectory is itself a case study in implementation fragility: the provisional

---

79 BRASIL. Câmara dos Deputados. PL nº 2.338, de 2023 — ficha de tramitação. Brasília, 2026. Disponível em: https://www.camara.leg.br/proposicoesWeb/fichadetramitacao?idProposicao=2487262. Acesso em: 31 jul. 2026; DIAP. PL 2338/23: votação do projeto sobre inteligência artificial está prevista apenas para dezembro. Brasília, 2026.

80 BIONI, Bruno; SCHERTEL MENDES, Laura; ALMEIDA, Virgilio. Brasil não deve copiar Europa ou Texas para regular IA. *Folha de S.Paulo*, São Paulo, 21 jul. 2025. Disponível em: https://gesel.ie.ufrj.br/wp-content/uploads/2025/07/Artigo-Bruno-Bioni-Laura-Schertel-Mendes-e-Virgilio-Almeida-21.07.2025.pdf. Acesso em: 31 jul. 2026.

81 BRASIL. Ministério da Ciência, Tecnologia e Inovação (MCTI); CGEE. *Plano Brasileiro de Inteligência Artificial (PBIA) 2024-2028*. Brasília: MCTI, 2024.

measure that anticipated the incentives (MP No. 1,318/2025) lapsed without a Senate vote in February 2026, leaving the incentive framework dependent on the pending bill.[82] More fundamentally, critics, including Brazil's leading scientific societies and digital-rights organizations, note that fiscal incentives for hyperscaler data centres, without conditions of national control, jurisdiction or data governance, risk subsidizing the very foreign-controlled infrastructure whose risks the preceding sections have mapped, deepening rather than reversing digital dependence.[83]

3.5 Why compute, not just data

For a volume on the regulation of AI, one further step is necessary: the analysis cannot stop at data. The distinction is simple: data is the information that systems store and learn from; compute is the processing capacity, the chips and cloud machines, needed to train and run them. The governance-of-compute literature has established that computing power is the most governable input to advanced AI, because it is physical, quantifiable, excludable and concentrated in an extremely short supply chain, and that cloud providers function as natural regulatory intermediaries: whoever controls compute can monitor, condition and ultimately deny the capacity to train and deploy the most advanced AI systems.[84] This turns our argu-

---

[82] BRASIL. Senado Federal. PL nº 278, de 2026 (REDATA). Brasília, 2026. Disponível em: https://www25.senado.leg.br/web/atividade/materias/-/materia/172786. Acesso em: 31 jul. 2026; SENADO FEDERAL. MP do Redata perde validade e incentivos dependem de projeto de lei. Brasília, 27 fev. 2026. Disponível em: https://www12.senado.leg.br/noticias/materias/2026/02/27/mp-do-redata-perde-validade-e-incentivos-dependem-de-projeto-de-lei. Acesso em: 1 ago. 2026.

[83] DATA PRIVACY BRASIL. Nota técnica sobre o PL que institui o Regime Especial de Tributação para Serviços de Datacenter — REDATA. São Paulo, 2026. Disponível em: https://www.dataprivacybr.org/data-privacy-brasil-publica-nota-tecnica-sobre-o-pl-que-institui-o-regime-especial-de-tributacao-para-servicos-de-datacenter-redata/. Acesso em: 31 jul. 2026. For a quantitative measure of Brazil's position, see SILVEIRA, Sergio Amadeu da; XIONG, Jeff. Índice de soberania digital. *Liinc em Revista*, v. 21, n. 1, e7451, 2025. DOI: 10.18617/liinc.v21i1.7451.

[84] SASTRY, Girish et al. *Computing power and the governance of artificial intelligence*. arXiv:2402.08797, 2024. DOI: 10.48550/arXiv.2402.08797; HEIM, Lennart et al. *Governing through the cloud: the intermediary role of compute providers in AI regulation*. arXiv:2403.08501, 2024. DOI: 10.48550/arXiv.2403.08501.

ment from analogy into identity. The same corporate-controlled cloud infrastructure that exposes Brazilian data to extraterritorial legal access is the infrastructure on which Brazilian AI, public and private, will run. An AI regulatory framework that assigns obligations by risk category while ignoring who controls the compute on which high-risk systems run, and which foreign authorities can reach the data that trains and feeds them, would transpose to AI the very pattern of formal protection and material vulnerability documented in this section.

A brief comparative note is instructive here. Canada, a wealthy G7 country closely allied to the United States, has arrived at analogous concerns. Its Sovereign AI Compute Strategy, launched on 5 December 2024 with up to CAD 2 billion, is expressly justified by the need for domestic compute capacity under national control,[85] after studies showed most Canadian AI workloads running on U.S.-controlled infrastructure.[86] Canada's first attempt at AI legislation (the Artificial Intelligence and Data Act, part of Bill C-27) nonetheless died on the Order Paper when Parliament was prorogued on 6 January 2025, for domestic political reasons rather than under external pressure, leaving the country, like Brazil, with sovereign infrastructure ambitions and an unfinished regulatory framework. The lesson runs in both directions. Even for a country far richer and better integrated into U.S. supply chains, compute and data sovereignty has become a mainstream policy objective; and even there, the regulatory dimension lags behind the infrastructural one. Digital sovereignty is neither a Global South idiosyncrasy nor a rhetorical flourish. It has become a normal object of statecraft.[87]

CONCLUSION

This chapter has investigated how the CLOUD Act gives effect to U.S. technological hegemony over Brazil and has examined institutional pathways for reducing this

---

[85] CANADA. Innovation, Science and Economic Development Canada (ISED). *Canadian Sovereign AI Compute Strategy*. Ottawa: ISED, 2024.

[86] DOBBS, Graham; HIRSCH-ALLEN, Jake. *Can Canada compute? Policy options to close Canada's AI compute gap*. Toronto: The Dais, Toronto Metropolitan University, mar. 2024.

[87] SCASSA, Teresa. Sovereignty and the governance of artificial intelligence. *UCLA Law Review Discourse*, v. 71, p. 214-229, 2024.

vulnerability. Brazilian dependence on foreign infrastructure is structural and growing, and the cases analyzed evidence national fragility before foreign governments and digital criminals alike. The analysis of the CLOUD Act's architecture, reinforced by FISA Section 702 in its expanded post-2024 form, demonstrated its incompatibilities with fundamental principles of the LGPD and the absence of both legal and contractual protections. The events of 2025–2026 removed any remaining doubt that extraterritorial legal power is exercised rather than merely held in reserve: Brazil's own moves toward digital sovereignty have themselves drawn sanctions and trade retaliation. Our thesis should be stated with precision. This dependence is neither technologically inevitable nor quickly reversible. It is path-dependent, the product of accumulated policy choices, and it can be reduced incrementally, for selected critical public-sector workloads, through procurement discipline, technical controls and domestic capability investment, provided the missing premise is confronted honestly: substitution has capability and cost limits, and no current Brazilian infrastructure replaces hyperscaler scale.

The federal government's sovereign cloud initiative, although important, does not fully resolve the existing problems, and the state still spends more on foreign technology services than on building its own infrastructure. The legislature, through Bill No. 2,338/2023, remains silent on concrete measures to guarantee Brazilian digital sovereignty, while the REDATA regime risks subsidizing foreign-controlled infrastructure without sovereignty conditions. In the face of this diagnosis, we propose an agenda based on four articulated axes. In the vocabulary of section 1, the agenda is a counter-hegemonic strategy, a war of position waged on the terrain of law, procurement and infrastructure. It does not imagine an exit from the global digital economy; it contests, incrementally, the terms of Brazil's insertion in it.[88]

Three objections to any such agenda must be confronted before its content is stated, because they are the objections a serious interlocutor will raise. First, the authoritarian-cover objection: state-sovereignty vocabulary in cyberspace has served authoritarian consolidation elsewhere, and its most influential critic argues that the sovereignty frame should be rejected outright.[89] The answer is institutional design.

---

[88] COX, cit.; GILL; LAW, cit.

[89] MUELLER, Milton L. Against sovereignty in cyberspace. *International Studies Review*, v. 22, n.

Everything proposed below is rights-based, judicially reviewable and transparent, anchored in a constitutional order that has made data protection a fundamental right; that is the opposite of the executive-controlled, opaque model the objection targets. Second, the cost objection: data-flow restrictions measurably reduce productivity, and localization mandates can raise costs without improving security.[90] The answer is scope. The agenda is sectoral and conditional, confined to critical public-sector workloads; it is not a general localization mandate for the economy. Third, the zero-risk fallacy: no lawful transfer regime can guarantee the absence of foreign-government access, and a policy that demands immunity will fail by its own standard.[91] The answer lies in the objective itself: not immunity, which no regime can deliver, but a proportionate reduction of risk for defined classes of data. With these boundaries drawn, the four axes follow.

First, the creation of a Brazilian blocking statute, better described as a sectoral foreign-access statute, inspired by the function Article 48 performs in the GDPR, as consolidated by the EDPB's Guidelines 02/2024, but adapted to national specificities and designed for enforceability rather than symbolism. Its constitutional foundation lies in national sovereignty (Article 1, I, of the 1988 Federal Constitution), the dignity of the human person (Article 1, III), the protection of personal data as an autonomous fundamental right, recognized by the Supreme Federal Tribunal in the ADI 6,387 line of cases and constitutionalized by Amendment No. 115/2022 (Article 5, LXXIX), and the Union's competence to legislate on the protection and processing of personal data (Article 22, XXX).[92]

Enforceability requires three articulated layers. The record of the EU Blocking

4, p. 779-801, 2020. DOI: 10.1093/isr/viz044.

90 FERRACANE, Martina F.; KREN, Janez; VAN DER MAREL, Erik. Do data policy restrictions impact the productivity performance of firms and industries? *Review of International Economics*, v. 28, n. 3, p. 676-722, 2020. DOI: 10.1111/roie.12467.

91 CHRISTAKIS, Théodore. The "zero risk" fallacy: international data transfers, foreign governments' access to data and the need for a risk-based approach. SSRN, 20 Feb. 2024. Disponível em: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4732294. Acesso em: 1 ago. 2026.

92 BRASIL. Constituição da República Federativa do Brasil de 1988; BRASIL. Emenda Constitucional nº 115, de 10 de fevereiro de 2022; BRASIL. Supremo Tribunal Federal. ADI 6.387 MC-Ref. Rel. Min. Rosa Weber, j. 7 maio 2020.

Regulation reinforces the point: statutes enacted mainly as signals go under-enforced, because firms rationally comply with U.S. law even at the price of foreign penalties.[93] The Brazilian design therefore needs enforcement at three levels. The first layer is a judicial gateway. For a narrowly defined class of critical public-sector data (judicial, health, social security, national security), transfer, access or disclosure pursuant to a foreign order would require prior authorization by a Brazilian court, upon an analysis of proportionality, reciprocity and equivalent protection, with foreign requests channelled through instruments of international cooperation and with data subjects guaranteed adversarial proceedings and full defence. The second layer is a duty on controllers and processors under the LGPD, supervised and sanctioned by the ANPD within its statutory remit, including fines proportional to revenue in Brazil and joint liability for intermediaries that facilitate unauthorized access, with any contractual clause waiving the protection deemed null and void. The third layer is ex ante procurement conditionality. Within the framework of the Public Procurement Law (Law No. 14,133/2021), which permits objective technical and security requirements tied to the object of the contract, though not blanket nationality restrictions, public contracts for critical workloads should require customer-held keys, auditability, portability and exit plans, local incident response, and contractual notification duties regarding foreign legal process.[94] France's blocking statute (Law No. 68-678, as reinforced in the Sapin II era) and the EU Blocking Regulation offer precedents for this design, and above all cautionary lessons on enforcement.

Second, a substantial and phased expansion of the sovereign cloud project. Rather than a rhetorical target of total migration, the goal should be the classification of

---

93 EUROPEAN UNION. Council Regulation (EC) No 2271/96 of 22 November 1996 protecting against the effects of the extra-territorial application of legislation adopted by a third country, and actions based thereon or resulting therefrom. *Official Journal of the European Communities*, L 309, p. 1-6, 29 nov. 1996; FRANCE. Loi nº 68-678 du 26 juillet 1968 relative à la communication de documents et renseignements d'ordre économique, commercial, industriel, financier ou technique à des personnes physiques ou morales étrangères (as amended by Loi nº 80-538 du 16 juillet 1980). On the under-enforcement diagnosis, see CHRISTAKIS, cit.

94 BRASIL. Lei nº 14.133, de 1º de abril de 2021 (Lei de Licitações e Contratos Administrativos), arts. 5º, 9º e 26; AGÊNCIA NACIONAL DE PROTEÇÃO DE DADOS (ANPD). Resolução CD/ANPD nº 19, de 23 de agosto de 2024 (Regulamento de Transferência Internacional de Dados Pessoais). Brasília: ANPD, 2024.

government workloads by criticality, with sensitive systems (health, social security, defence and justice) migrated first under explicit timelines, total-cost-of-ownership assessment, interoperability and open-standards requirements, free and open-source software where feasible, periodic audit and exit tests, and review clauses. The control hierarchy described in section 3.4 should govern: customer-held keys and confidential computing where those controls suffice; sovereign infrastructure, or disconnected operation, for the residual highest-risk tier. Public incentives for data centres, including the PBIA's sovereign infrastructure axis and the fiscal instruments of REDATA, should carry conditions of national jurisdiction, data governance, key custody, audit rights and public purpose, so that the state subsidizes sovereignty rather than dependence. The Huawei lesson should be generalized: supplier diversification is not sovereignty; control is.

Third, the amendment of Bill No. 2,338/2023, whose prolonged passage through the Chamber of Deputies keeps this window open, to provide that high-risk AI systems, when used by the public authorities or directed at the provision of public services, must host and process their data on infrastructure located on national territory, controlled by Brazilian public or private entities and subject to national legislation, or protected by the technical controls of the first tier of the hierarchy above. The rationale is the one developed in section 3.5. Compute is the governable input of AI, and cloud providers are the regulatory intermediaries; an AI framework that regulates risk categories while ignoring infrastructure control would reproduce, for AI, the pattern of formal protection and material vulnerability this text has described.[95]

Fourth, and finally, a bilateral track, stated with precision and sequenced last. The instrument in question would be an executive agreement under 18 U.S.C. § 2523, the mechanism through which the United Kingdom (2019) and Australia (2021) obtained qualifying-government status.[96] Honesty requires acknowledging the asym-

---

[95] SASTRY et al., cit.; HEIM et al., cit.

[96] UNITED KINGDOM; UNITED STATES OF AMERICA. *Agreement on Access to Electronic Data for the Purpose of Countering Serious Crime*. Washington, DC, 3 Oct. 2019; UNITED STATES. 18 U.S.C. § 2523. Disponível em: https://www.law.cornell.edu/uscode/text/18/2523. Acesso em: 1 ago. 2026.

metry. Certification depends on the U.S. Attorney General attesting that the partner's law satisfies U.S.-defined criteria of privacy and civil liberties, a test whose application to Brazil would be contested, and a negotiation the current conjuncture renders particularly unbalanced. Two considerations nonetheless keep the axis on the agenda. Brazil has its own demand-side interest: Brazilian courts and prosecutors are themselves substantial requesters of data held by U.S. providers, as the record of the STF's ADC 51 proceedings documents, an interest in faster, rights-respecting reciprocal access rather than merely in blocking.[97] And the first three axes strengthen this fourth rather than undermining it. A functioning domestic gateway and demonstrated technical control are bargaining assets; the blocking statute is the bargaining position of which the bilateral agreement is a possible negotiated outcome. Any such agreement must contain robust safeguards: prior notification of the Brazilian government, judicial review by a Brazilian court, prohibition of mass surveillance, and reciprocity that does not expose national data to U.S. authorities without protections equivalent to those enjoyed by U.S. citizens. Because the conditions for such a negotiation do not currently obtain, the first three axes, which depend only on Brazilian institutions, should not wait for the fourth.

The implementation of this agenda would face obstacles that cannot be underestimated, from material, political and economic constraints to private interests acting through lobbying in the National Congress, and, now, to external pressure that explicitly targets sovereignty-enhancing measures. These limitations do not justify inaction. The costs of digital subordination are high and lasting. The question is not whether Brazil can achieve complete technological autonomy, which no country can, but what margin of decision it will retain over questions bound up with the fundamental rights of its citizens and strategic national interests.

We have sought to demonstrate Brazilian vulnerability before the U.S. CLOUD Act as one piece of an architecture that combines the economic domination of big tech, characteristic of surveillance capitalism, with the legal-political domination of

---

97 BRASIL. Supremo Tribunal Federal. ADC 51 — audiência pública, 10 fev. 2020 (registro de que o Brasil figurava entre os principais países requerentes de dados a provedores estadunidenses). Disponível em: https://portal.stf.jus.br/noticias/verNoticiaDetalhe.asp?idConteudo=436651&tip=UN. Acesso em: 1 ago. 2026.

the United States in the dispute over global technological hegemony, a dispute in which AI has raised the stakes, since whoever controls compute, data and models will shape the infrastructures of decision-making itself. As Pontin observed in the debate that helped motivate this chapter, *"artificial intelligence is producing text, and in producing text, it produces reality, and in producing reality, it produces hegemony"*.[98] Brazil's dependence on foreign infrastructure, the face of digital colonialism, results from political choices that can be revisited, not reversed overnight but redirected, through legislative action, public investment and the construction of alternatives. Digital sovereignty will never be total. Pursued as a disciplined counter-hegemonic strategy, it can be significantly greater than it is today.

REFERENCES

AGÊNCIA BRASIL. "Verdade prevaleceu", diz Moraes sobre fim de sanções. Brasília, dez. 2025. Disponível em: https://agenciabrasil.ebc.com.br/justica/noticia/2025-12/verdade-prevaleceu-diz-moraes-sobre-fim-de-sancoes. Acesso em: 31 jul. 2026.

AGÊNCIA BRASIL. Sites e aplicativo do Ministério da Saúde sofrem ataque cibernético. 10 dez. 2021. Disponível em: https://agenciabrasil.ebc.com.br/saude/noticia/2021-12/sites-e-aplicativo-do-ministerio-da-saude-sofrem-ataque-cibernetico. Acesso em: 27 dez. 2025.

AGÊNCIA GOV. Soberania digital: Brasil consolida a Nuvem de Governo como passo para a independência e a proteção de dados. Brasília, set. 2025. Disponível em: https://agenciagov.ebc.com.br/noticias/202509/na-soberania-digital-brasil-consolida-a-nuvem-de-governo-como-passo-para-a-independencia-e-a-protecao-de-dados. Acesso em: 31 jul. 2026.

AGÊNCIA NACIONAL DE PROTEÇÃO DE DADOS (ANPD). Resolução CD/ANPD nº 19, de 23 de agosto de 2024 (Regulamento de Transferência Internacional de Dados Pessoais). Brasília: ANPD, 2024.

BAGUETE. Ministério da Saúde migra para AWS com Embratel. 22 abr. 2020. Disponível em: https://www.baguete.com.br/noticias/22/04/2020/ministerio-da-saude-migra-para-aws-com-embratel. Acesso em: 31 dez. 2025.

BIONI, Bruno; SCHERTEL MENDES, Laura; ALMEIDA, Virgilio. Brasil não deve copiar Europa ou Texas para regular IA. *Folha de S.Paulo*, São Paulo, 21 jul. 2025. Disponível em: https://gesel.ie.ufrj.br/wp-content/uploads/2025/07/Artigo-Bruno-Bioni-Laura-Schertel-Mendes-e-Virgilio-Almeida-21.07.2025.pdf. Acesso em: 31 jul. 2026.

[98] PONTIN, cit., a 1h24min19s; tradução nossa.

BORGOGNO, Oscar; SAVINI ZANGRANDI, Michele. Chinese data governance and trade policy: from cyber sovereignty to the quest for digital hegemony? *Journal of Contemporary China*, v. 33, n. 148, p. 578-602, 2024.

BRADFORD, Anu. *Digital empires: the global battle to regulate technology*. New York: Oxford University Press, 2023.

BRASIL DE FATO. ConecteSUS fora do ar: governo contratou nuvem privada sem serviço de segurança. São Paulo, 21 dez. 2021. Disponível em: https://www.brasildefato.com.br/2021/12/21/conectesus-fora-do-ar-governo-contratou-nuvem-privada-sem-servico-de-seguranca. Acesso em: 31 dez. 2025.

BRASIL. Constituição da República Federativa do Brasil de 1988. Brasília: Presidência da República, 1988.

BRASIL. Câmara dos Deputados. Projeto de Lei nº 2.338, de 2023 — ficha de tramitação. Brasília: Câmara dos Deputados, 2026. Disponível em: https://www.camara.leg.br/proposicoesWeb/fichadetramitacao?idProposicao=2487262. Acesso em: 31 jul. 2026.

BRASIL. Câmara dos Deputados. Resposta do SERPRO ao Requerimento de Informação nº 108/2025. Brasília, 2025. Disponível em: https://www.camara.leg.br/proposicoesWeb/prop_mostrarintegra?codteor=2888474&filename=Tramitacao-18-RIC-108-2025. Acesso em: 1 ago. 2026.

BRASIL. Emenda Constitucional nº 115, de 10 de fevereiro de 2022. Brasília: Presidência da República, 2022.

BRASIL. Lei nº 13.709, de 14 de agosto de 2018 (Lei Geral de Proteção de Dados Pessoais — LGPD). Brasília: Presidência da República, 2018.

BRASIL. Lei nº 14.133, de 1º de abril de 2021 (Lei de Licitações e Contratos Administrativos). Brasília: Presidência da República, 2021.

BRASIL. Ministério da Ciência, Tecnologia e Inovação (MCTI); CGEE. *Plano Brasileiro de Inteligência Artificial (PBIA) 2024-2028*. Brasília: MCTI, 2024.

BRASIL. Ministério da Gestão e da Inovação em Serviços Públicos (MGI). Gestão assina acordos com Serpro e Dataprev para criação de Nuvem de Governo. Brasília, 12 nov. 2024. Disponível em: https://www.gov.br/gestao/pt-br/assuntos/noticias/2024/novembro/gestao-assina-acordos-com-serpro-e-dataprev-para-criacao-de-nuvem-de-governo. Acesso em: 31 dez. 2025.

BRASIL. Ministério da Gestão e da Inovação em Serviços Públicos (MGI). Gestão, Serpro e Dataprev lançam serviços de nuvem de governo para órgãos federais. Brasília, 10 jun. 2025. Disponível em: https://www.gov.br/governodigital/pt-br/noticias/gestao-serpro-e-dataprev-lancam-servicos-de-nuvem-de-governo-para-orgaos-federais. Acesso em: 31 jul. 2026.

BRASIL. Secretaria de Governo Digital (SGD). O que é a diretriz Cloud First da SGD para o SISP? Brasília. Disponível em: https://www.gov.br/governodigital/pt-br/estrategias-e-governanca-digital/estrategias-e-politicas-digitais/computacao-em-nuvem/o-que-e-a-diretriz-cloud-first-

da-sgd-para-o-sisp. Acesso em: 1 ago. 2026.

BRASIL. Senado Federal. Projeto de Lei nº 278, de 2026 (REDATA — Regime Especial de Tributação para Serviços de Datacenter). Brasília: Senado Federal, 2026. Disponível em: https://www25.senado.leg.br/web/atividade/materias/-/materia/172786. Acesso em: 31 jul. 2026.

BRASIL. Supremo Tribunal Federal (STF). ADC 51 — audiência pública. Brasília, 10 fev. 2020. Disponível em: https://portal.stf.jus.br/noticias/verNoticiaDetalhe.asp?idConteudo=436651&tip=UN. Acesso em: 1 ago. 2026.

BRASIL. Supremo Tribunal Federal (STF). ADI 6.387 MC-Ref. Rel. Min. Rosa Weber, j. 7 maio 2020. Brasília: STF, 2020.

BRENNAN CENTER FOR JUSTICE. *Section 702 of FISA explained*. New York: Brennan Center for Justice, 2026. Disponível em: https://www.brennancenter.org/our-work/research-reports/section-702-fisa-explained. Acesso em: 31 jul. 2026.

BROEDERS, Dennis; CRISTIANO, Fabio; KAMINSKA, Monica. In search of digital sovereignty and strategic autonomy: normative power Europe to the test of its geopolitical ambitions. *JCMS: Journal of Common Market Studies*, v. 61, n. 5, p. 1261-1280, 2023. DOI: 10.1111/jcms.13462.

CANADA. Innovation, Science and Economic Development Canada (ISED). *Canadian Sovereign AI Compute Strategy*. Ottawa: ISED, 2024.

CARROLL, Stephanie Russo et al. The CARE Principles for Indigenous data governance. *Data Science Journal*, v. 19, art. 43, p. 1-12, 2020. DOI: 10.5334/dsj-2020-043.

CHINA. Data Security Law of the People's Republic of China. Beijing: National People's Congress, 2021. Disponível em: http://www.npc.gov.cn/englishnpc/c2759/c23934/202112/t20211209_385109.html. Acesso em: 1 ago. 2026.

CHINA. National Intelligence Law of the People's Republic of China. Beijing: National People's Congress, 2017 (as amended, 2018).

CHRISTAKIS, Théodore. The "zero risk" fallacy: international data transfers, foreign governments' access to data and the need for a risk-based approach. SSRN, 20 Feb. 2024. Disponível em: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4732294. Acesso em: 1 ago. 2026.

COCHRANE, Tim. Hiding in the eye of the storm cloud: how CLOUD Act agreements expand U.S. extraterritorial investigatory powers. *Duke Journal of Comparative & International Law*, v. 32, n. 1, p. 47-92, 2021.

COHEN, Julie E. Review of Zuboff's The Age of Surveillance Capitalism. *Surveillance & Society*, v. 17, n. 1/2, p. 240-245, 2019. DOI: 10.24908/ss.v17i1/2.13144.

COLI, Anna Luiza; AVRITZER, Gabriel Caetano. Da promessa da soberania digital à realidade do aluguel tecnológico. Belo Horizonte: UFMG, 5 set. 2025. Disponível em: https://www.ufmg.br/comunicacao/noticias/opiniao/da-promessa-da-soberania-digital-a-realidade-do-aluguel-tecnologico/. Acesso em: 31 dez. 2025.

CONGRESSIONAL RESEARCH SERVICE (CRS). *FISA Section 702 and the 2024 RISAA*. CRS Report R48592. Washington, DC: Library of Congress, 2025. Disponível em: https://crsreports.congress.gov/product/pdf/R/R48592. Acesso em: 31 jul. 2026.

CONGRESSIONAL RESEARCH SERVICE (CRS). Law enforcement access to overseas data under the CLOUD Act. LSB10125. Washington, DC: Library of Congress. Disponível em: https://www.congress.gov/crs-product/LSB10125. Acesso em: 1 ago. 2026.

CONSULTOR JURÍDICO. CNJ suspende contrato de R$ 1,3 bi entre TJ-SP e Microsoft. 21 fev. 2019. Disponível em: https://www.conjur.com.br/2019-fev-21/cnj-suspende-contrato-13-bi-tj-sp-microsoft. Acesso em: 27 dez. 2025.

CONSULTOR JURÍDICO. Lei Magnitsky projeta ordem jurídica dos EUA em território brasileiro. São Paulo, 31 out. 2025. Disponível em: https://www.conjur.com.br/2025-out-31/lei-magnitsky-projeta-ordem-juridica-dos-eua-em-territorio-brasileiro/. Acesso em: 31 jul. 2026.

CONSULTOR JURÍDICO. TJ-SP rescinde contrato de R$ 1,3 bilhão com a Microsoft. 20 maio 2020. Disponível em: https://www.conjur.com.br/2020-mai-20/tj-sp-rescinde-contrato-13-bilhao-microsoft/. Acesso em: 1 ago. 2026.

COULDRY, Nick; MEJIAS, Ulises A. Data colonialism: rethinking big data's relation to the contemporary subject. *Television & New Media*, v. 20, n. 4, p. 336-349, 2019.

COULDRY, Nick; MEJIAS, Ulises A. *The costs of connection: how data is colonizing human life and appropriating it for capitalism*. Stanford: Stanford University Press, 2019.

COUTURE, Stéphane; TOUPIN, Sophie. What does the notion of "sovereignty" mean when referring to the digital? *New Media & Society*, v. 21, n. 10, p. 2305-2322, 2019.

COX, Robert W. Gramsci, hegemony and international relations: an essay in method. *Millennium: Journal of International Studies*, v. 12, n. 2, p. 162-175, 1983. DOI: 10.1177/03058298830120020701. (Reprinted in: COX, Robert W.; SINCLAIR, Timothy J. *Approaches to world order*. Cambridge: Cambridge University Press, 1996. p. 124-143.)

DATA PRIVACY BRASIL. Nota técnica sobre o PL que institui o Regime Especial de Tributação para Serviços de Datacenter — REDATA. São Paulo: Data Privacy Brasil, 2026. Disponível em: https://www.dataprivacybr.org/data-privacy-brasil-publica-nota-tecnica-sobre-o-pl-que-institui-o-regime-especial-de-tributacao-para-servicos-de-datacenter-redata/. Acesso em: 31 jul. 2026.

DIAP — DEPARTAMENTO INTERSINDICAL DE ASSESSORIA PARLAMENTAR. PL 2338/23: votação do projeto sobre inteligência artificial está prevista apenas para dezembro. Brasília: DIAP, 2026. Disponível em: https://www.diap.org.br/index.php/noticias/noticias/92249. Acesso em: 31 jul. 2026.

DOBBS, Graham; HIRSCH-ALLEN, Jake. *Can Canada compute? Policy options to close Canada's AI compute gap*. Toronto: The Dais, Toronto Metropolitan University, mar. 2024.

ERIE, Matthew S.; STREINZ, Thomas. The Beijing Effect: China's Digital Silk Road as transnational data governance. *New York University Journal of International Law and Politics*, v. 54, n. 1, p. 1-91, 2021.

EUROPEAN DATA PROTECTION BOARD (EDPB). *Guidelines 02/2024 on Article 48 GDPR*. Version 2.0, adopted 4 June 2025; Version 2.1, 20 June 2025. Brussels: EDPB, 2025. Disponível em: https://www.edpb.europa.eu/system/files/2025-06/edpb_guidelines_202402_article48_v2_en.pdf. Acesso em: 31 jul. 2026.

EUROPEAN UNION. Council Regulation (EC) No 2271/96 of 22 November 1996 protecting against the effects of the extra-territorial application of legislation adopted by a third country, and actions based thereon or resulting therefrom. *Official Journal of the European Communities*, L 309, p. 1-6, 29 nov. 1996.

FARRELL, Henry; NEWMAN, Abraham L. *Underground empire: how America weaponized the world economy*. New York: Henry Holt, 2023.

FARRELL, Henry; NEWMAN, Abraham L. Weaponized interdependence: how global economic networks shape state coercion. *International Security*, v. 44, n. 1, p. 42-79, 2019. DOI: 10.1162/isec_a_00351.

FERRACANE, Martina F.; KREN, Janez; VAN DER MAREL, Erik. Do data policy restrictions impact the productivity performance of firms and industries? *Review of International Economics*, v. 28, n. 3, p. 676-722, 2020. DOI: 10.1111/roie.12467.

FRANCE. Loi nº 68-678 du 26 juillet 1968 relative à la communication de documents et renseignements d'ordre économique, commercial, industriel, financier ou technique à des personnes physiques ou morales étrangères (as amended by Loi nº 80-538 du 16 juillet 1980).

FRANCE. Sénat. Commission d'enquête sur les coûts et les modalités effectifs de la commande publique et la mesure de son effet d'entraînement sur l'économie française. Audition de MM. Anton Carniaux et Pierre Lagarde, Microsoft France. Séance du mardi 10 juin 2025. Disponível em: https://www.senat.fr/compte-rendu-commissions/20250609/ce_commande_publique.html. Acesso em: 1 ago. 2026.

FRATINI, Samuele; HINE, Emmie; NOVELLI, Claudio; ROBERTS, Huw; FLORIDI, Luciano. Digital sovereignty: a descriptive analysis and a critical evaluation of existing models. *Digital Society*, v. 3, n. 3, art. 59, 2024. DOI: 10.1007/s44206-024-00146-7.

FURTADO, Renato Guimarães; CUNHA, Simone Evangelista. Inteligência artificial, data centers e colonialismo digital: impactos socioambientais e geopolíticos a partir do Sul Global. *Liinc em Revista*, v. 20, n. 2, 2024.

GERMAIN, Randall D.; KENNY, Michael. Engaging Gramsci: international relations theory and the new Gramscians. *Review of International Studies*, v. 24, n. 1, p. 3-21, 1998. DOI: 10.1017/S0260210598000035.

GILL, Stephen; LAW, David. Global hegemony and the structural power of capital. *International Studies Quarterly*, v. 33, n. 4, p. 475-499, 1989. DOI: 10.2307/2600523.

GLASZE, Georg et al. Contested spatialities of digital sovereignty. *Geopolitics*, v. 28, n. 2, p. 919-958, 2022. DOI: 10.1080/14650045.2022.2050070.

GREENWALD, Glenn; MACASKILL, Ewen. NSA Prism program taps in to user data of Apple, Google and others. *The Guardian*, London, 6 jun. 2013. Disponível em: https://www.theguardian.com/world/2013/jun/06/us-tech-giants-nsa-data. Acesso em: 3 ago. 2026.

HEIM, Lennart et al. *Governing through the cloud: the intermediary role of compute providers in AI regulation*. arXiv:2403.08501, 2024. DOI: 10.48550/arXiv.2403.08501.

IVES, Peter; SHORT, Nicola. On Gramsci and the international: a textual analysis. *Review of International Studies*, v. 39, n. 3, p. 621-642, 2013. DOI: 10.1017/S0260210512000526.

KWET, Michael. Digital colonialism: US empire and the new imperialism in the Global South. *Race & Class*, v. 60, n. 4, p. 3-26, 2019. DOI: 10.1177/0306396818823172.

MASSUCHIN, Michele Goulart; TAVARES, Camila Quesada; COMEL, Naiza. Soberania digital e informacional: desafios da gestão de tecnologias, da comunicação pública e os reflexos para a cidadania. In: ENCONTRO ANUAL DA COMPÓS, 34., 2025, Curitiba. Anais [...]. Curitiba: UFPR, 2025.

McCARTHY, Daniel R. *Power, information technology, and international relations theory: the power and politics of US foreign policy and the internet*. Basingstoke: Palgrave Macmillan, 2015.

McCARTHY, Daniel R. The meaning of materiality: reconsidering the materialism of Gramscian IR. *Review of International Studies*, v. 37, n. 3, p. 1215-1234, 2011. DOI: 10.1017/S026021051000077X.

MICHELS, Johan David; MILLARD, Christopher; WALDEN, Ian; WUERMELING, Ulrich. *Cloud sovereignty and the GDPR, Part One: US government access to European data*. Queen Mary Law Research Paper No. 431/2024. London: Queen Mary University of London, 2024. Disponível em: https://ssrn.com/abstract=4911552. Acesso em: 27 dez. 2025.

MICROSOFT. Controls and principles in sovereign public cloud. Redmond, 2026. Disponível em: https://learn.microsoft.com/en-us/azure/azure-sovereign-clouds/public/overview-controls-principles. Acesso em: 1 ago. 2026.

MICROSOFT. European digital commitments. Redmond, 30 Apr. 2025. Disponível em: https://blogs.microsoft.com/on-the-issues/2025/04/30/european-digital-commitments/. Acesso em: 1 ago. 2026.

MIGALHAS. CNJ mantém suspensa contratação do TJ-SP com Microsoft, mas permite estudos técnicos. 9 abr. 2019. Disponível em: https://www.migalhas.com.br/quentes/300031/cnj-mantem-suspensa-contratacao-do-tj-sp-com-microsoft--mas-permite-estudos-tecnicos. Acesso em: 1 ago. 2026.

MIGALHAS. EUA revogam sanções da lei Magnitsky contra Moraes e esposa. dez. 2025. Disponível em: https://www.migalhas.com.br/quentes/446354/eua-revogam-sancoes-da-lei-magnitsky-contra-moraes-e-esposa. Acesso em: 31 jul. 2026.

MOBILE TIME. Serpro: nuvem soberana hospeda 5 mil sistemas do governo e 50 terabytes. 11 maio 2026. Disponível em: https://www.mobiletime.com.br/noticias/11/05/2026/serpro-nuvem-soberana/. Acesso em: 31 jul. 2026.

MOURA, Stéfano Mariotto de. "Economia política dos dados" como alternativa à "governança da internet": uma perspectiva desde o sul global. *Brazilian Journal of International Relations*, v. 14, e025021, 2025. DOI: 10.36311/2237-7743.2025.v14.e025021.

MUELLER, Milton L. Against sovereignty in cyberspace. *International Studies Review*, v. 22, n. 4, p. 779-801, 2020. DOI: 10.1093/isr/viz044.

POHLE, Julia; NANNI, Riccardo; SANTANIELLO, Mauro. *Unthinking digital sovereignty: a critical reflection on origins, objectives, and practices*. *Policy & Internet*, v. 16, n. 4, p. 666-671, 2024. DOI: 10.1002/poi3.437.

POHLE, Julia; THIEL, Thorsten. *Digital sovereignty. Internet Policy Review*, v. 9, n. 4, 2020. DOI: 10.14763/2020.4.1532.

POLIDO, Fabrício Bertini Pasquot. Estado, soberania digital e tecnologias emergentes: interações entre direito internacional, segurança cibernética e inteligência artificial. *Revista de Ciências do Estado*, v. 9, n. 1, 2024.

PONTIN, Fabrício. [Participação]. In: ABAL, Felipe Cittolin; DIVAN, Gabriel. #471 "A IA vai curar o câncer?". Viracasacas [Podcast], 17 fev. 2026. Disponível em: https://youtu.be/F9MlIgTkodE. Acesso em: 1 ago. 2026.

PPC LAND. Microsoft can't protect French data from US government access. 18 July 2025. Disponível em: https://ppc.land/microsoft-cant-protect-french-data-from-us-government-access/. Acesso em: 26 dez. 2025.

PRIVACY AND CIVIL LIBERTIES OVERSIGHT BOARD (PCLOB). *Report on the surveillance program operated pursuant to Section 702 of the Foreign Intelligence Surveillance Act*. Washington, DC: PCLOB, 2 jul. 2014.

PRIVACY AND CIVIL LIBERTIES OVERSIGHT BOARD (PCLOB). *Report on the surveillance program operated pursuant to Section 702 of the Foreign Intelligence Surveillance Act: staff report*. Washington, DC: PCLOB, 2 abr. 2026.

RIKAP, Cecilia. Amazon: a story of accumulation through intellectual rentiership and predation. *Competition & Change*, v. 26, n. 3-4, p. 436-466, 2022. DOI: 10.1177/1024529420932418.

SANTANIELLO, Mauro. *Attributes of digital sovereignty: a conceptual framework*. *Geopolitics*, v. 31, n. 2, p. 788-809, 2026 (publicado online em 2025). DOI: 10.1080/14650045.2025.2521548.

SASTRY, Girish et al. *Computing power and the governance of artificial intelligence*. arXiv:2402.08797,

2024. DOI: 10.48550/arXiv.2402.08797.

SCASSA, Teresa. Sovereignty and the governance of artificial intelligence. *UCLA Law Review Discourse*, v. 71, p. 214-229, 2024.

SENADO FEDERAL. MP do Redata perde validade e incentivos dependem de projeto de lei. Brasília, 27 fev. 2026. Disponível em: https://www12.senado.leg.br/noticias/materias/2026/02/27/mp-do-redata-perde-validade-e-incentivos-dependem-de-projeto-de-lei. Acesso em: 1 ago. 2026.

SERPRO. Nuvem Serpro e inteligência artificial nacional. Brasília, 31 jul. 2024. Disponível em: https://www.serpro.gov.br/menu/noticias/noticias-2024/nuvem-serpro-inteligencia-artificial-nacional. Acesso em: 1 ago. 2026.

SILVA, Ergon Cugler de Moraes; ROCHA, Isabela; VAZ, José Carlos; VENEZIANI, Julia Ribeiro de Almeida; MODANEZ, Camila de Camargo. *Contratos, códigos e controle: a influência das big techs no Estado brasileiro*. São Paulo: GETIP/EACH/USP; Brasília: GEPSI/IREL-UnB, jul. 2025. Disponível em: https://bit.ly/contratos-big-techs. Acesso em: 27 dez. 2025.

SILVEIRA, Sergio Amadeu da; XIONG, Jeff. Índice de soberania digital. *Liinc em Revista*, v. 21, n. 1, e7451, 2025. DOI: 10.18617/liinc.v21i1.7451.

STRANGE, Susan. *States and markets*. London: Pinter, 1988.

SWIRE, Peter. *When does GDPR act as a blocking statute: the relevance of a lawful basis for transfer*. Georgia Tech Scheller College of Business Research Paper, 2019. Disponível em: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3473187. Acesso em: 31 dez. 2025.

SYNERGY RESEARCH GROUP. Cloud market annual revenue run rate topped half a trillion dollars in Q1 as growth surge continues. Reno, May 2026. Disponível em: https://www.srgresearch.com/articles/cloud-market-annual-revenue-run-rate-topped-half-a-trillion-dollars-in-q1-as-growth-surge-continues. Acesso em: 1 ago. 2026.

THE REGISTER. Microsoft exec admits it 'cannot guarantee' data sovereignty. 25 July 2025. Disponível em: https://www.theregister.com/off-prem/2025/07/25/microsoft_admits_it_cannot_guarantee/. Acesso em: 31 jul. 2026.

UNITED KINGDOM; UNITED STATES OF AMERICA. *Agreement on Access to Electronic Data for the Purpose of Countering Serious Crime*. Washington, DC, 3 Oct. 2019.

UNITED STATES. 18 U.S.C. § 2523 (executive agreements on access to data by foreign governments). Disponível em: https://www.law.cornell.edu/uscode/text/18/2523. Acesso em: 1 ago. 2026.

UNITED STATES. 18 U.S.C. § 2703(h) (comity analysis of legal process seeking contents of wire or electronic communication). Disponível em: https://www.law.cornell.edu/uscode/text/18/2703. Acesso em: 1 ago. 2026.

UNITED STATES. *Clarifying Lawful Overseas Use of Data Act (CLOUD Act)*. 18 U.S.C. § 2713. Washington, DC: U.S. Congress, 2018.

UNITED STATES. Department of Justice. CLOUD Act resources. Washington, DC. Disponível em:

https://www.justice.gov/criminal/cloud-act-resources. Acesso em: 1 ago. 2026.

UNITED STATES. Executive Order of 30 July 2025 (additional tariffs on Brazilian goods). *Federal Register*, Washington, DC, 5 Aug. 2025. Disponível em: https://www.federalregister.gov/documents/full_text/xml/2025/08/05/2025-14896.xml. Acesso em: 1 ago. 2026.

UNITED STATES. National Security Agency (NSA); Cybersecurity and Infrastructure Security Agency (CISA). *Use secure cloud key management practices*: cybersecurity information sheet. Washington, DC: NSA; CISA, mar. 2024. Disponível em: https://media.defense.gov/2024/Mar/07/2003407858/-1/-1/0/CSI-CloudTop10-Key-Management.PDF. Acesso em: 3 ago. 2026.

UNITED STATES. Office of the United States Trade Representative (USTR). Initiation of Section 301 Investigation: Brazil's Acts, Policies, and Practices Related to Digital Trade and Electronic Payment Services. *Federal Register*, Washington, DC, 18 July 2025. Disponível em: https://www.federalregister.gov/documents/2025/07/18/2025-13498/initiation-of-section-301-investigation-brazils-acts-policies-and-practices-related-to-digital-trade. Acesso em: 31 jul. 2026.

UNITED STATES. Office of the United States Trade Representative (USTR). USTR Section 301 Action on Brazil's Unreasonable Acts, Policies, and Practices. Washington, DC, 14 July 2026. Disponível em: https://ustr.gov/about/policy-offices/press-office/press-releases/2026/july/ustr-section-301-action-brazils-unreasonable-acts-policies-and-practices. Acesso em: 1 ago. 2026.

UNITED STATES. Office of the United States Trade Representative (USTR). USTR Section 301 Determination on Brazil's Unreasonable Acts, Policies, and Practices. Washington, DC, 1 June 2026. Disponível em: https://ustr.gov/about/policy-offices/press-office/press-releases/2026/june/ustr-section-301-determination-brazils-unreasonable-acts-policies-and-practices. Acesso em: 31 jul. 2026.

ZUBOFF, Shoshana. *The age of surveillance capitalism: the fight for a human future at the new frontier of power*. New York: PublicAffairs, 2019.

FROM AISI TO AISO: BRAZIL ON THE SECOND WAVE OF AI GOVERNANCE INSTITUTIONS

Fabro Boaz Steibel

PhD Political Communication (Leeds/UK), Executive Director at The Institute for Technology and Society of Rio (ITS Rio) (2013–). G20/UNDP DPI consultant for Brazil's MGI (2024); UN Tech Envoy DPI Safeguards WG (2024); ANPD council member; OGP Independent Reporting Mechanism (3 terms). Berkman Klein/Harvard affiliated researcher.

Natan Jamil

LL.B. with honors (PUC-SP), Master's candidate in Philosophy and General Theory of Law, Universidade de São Paulo. Researcher in Law and Technology at ITS Rio and Research Assistant at the London School of Economics and Political Science (LSE). Former data protection lawyer at Mattos Filho Advogados.

Ana Carolina Britto

LL.B. (USP). Scheduled to begin her masters studies at the Commercial Law (USP) in the second semester of 2026. Jr. Researcher in Law and Technology at ITS Rio.

DOI: https://doi.org/10.59224/dti8.ch5

Abstract: Since 2023, AI governance has produced a recognisable institution: the AI Safety Institute (AISI). Built by jurisdictions upstream in the AI value chain, this first-wave of AISI centres on frontier-risk evaluation. This article argues that a new wave of AISI institutions is coming, and that the AISI's characteristics of the first

wave fits Brazil only partially. Brazil's most consequential questions concern deployment, access and adaptation, and not only misuse and safety alone. As a response to that, we propose the adoption of the model AISO - AI Safety and Opportunities Institute, preserving the technical, advisory and networked character of first-wave bodies while treating opportunity as a substantive mandate.

Keywords: AI Safety Institutes; AI governance; Brazil.

Table of contents: 1. Introduction. 2. AISIs and the First Wave. 2.1 What are AISIs. 2.2 The AISIs' Origin. 2.2.1 UK AI Safety Summit. 2.2.2 AI Seoul Summit. 2.2.3 AI Action Summit (Paris). 2.2.4 India AI Impact Summit. 2.3 AISIs - First wave. 2.3.1 United Kingdom. 2.3.2 United States of America. 2.3.3 Singapore. 2.3.4 Canada. 2.3.5 Japan. 2.3.6 France. 2.3.7 European Union. 2.3.8 South Korea. 3. From AISI to AISO: The Brazilian Proposal. 4. An AISO in Brazil: Characteristics and Functions in Practice. 4.1 Characteristics. 4.2 Functions. 5. Conclusion.

1. INTRODUCTION

The international governance of artificial intelligence (AI) has, since 2023, developed a recognisable institutional form: the AI Safety Institute (AISI). Established first in the United Kingdom (UK) and the United States (USA) on the margins of the Bletchley Summit[1], and subsequently replicated across Japan, Singapore, Canada, France, South Korea, and the European Union (EU), these bodies share a common profile: technical, advisory, embedded within government, and organised around the empirical evaluation of frontier AI systems. Their proliferation has been rapid

[1] ARAUJO, Renan; FORT, Kristina; GUEST, Oliver. *Understanding the First Wave of AI Safety Institutes: Characteristics, Functions, and Challenges*. Oxford: Institute for AI Policy and Strategy (IAPS), p. 7. Oct. 2024. Available at: https://doi.org/10.48550/arXiv.2410.09219. Accessed: 22 May 2026; CUÉLLAR, Mariano-Florentino. *The UK AI Safety Summit Opened a New Chapter in AI Diplomacy*. Washington, D.C.: Carnegie Endowment for International Peace. Nov. 2023. Available at: https://carnegieendowment.org/posts/2023/11/the-uk-ai-safety-summit-opened-a-new-chapter-in-ai-diplomacy. Accessed: 20 May 2026.

enough to constitute what the policy literature now calls a "first-wave"[2].

As the first-wave AISIs were built by and for jurisdictions that sit at the upstream end of the global AI value chain, this paper asks whether that same model is adequate for Brazil and, by extension, for the Global South[3]. Brazil occupies a structurally different position. It does not yet host firms developing frontier general-purpose models, its AI deployment landscape is characterised by dependence on infrastructures built elsewhere, and the policy questions that matter most to its population concern not only the misuse of frontier systems but the conditions under which AI can be productively developed and deployed domestically[4].

The thesis advanced here is that this asymmetry calls for a different institutional response. Rather than replicate an AISI, Brazil should establish an *AI Safety and Opportunities Institute* (AISO)[5]. An AISO preserves the technical, advisory, and internationally networked characteristics that give first-wave bodies their credibility, but adds a parallel and equally rigorous mandate: to evaluate how AI can be usefully

---

2 INSTITUTO DE TECNOLOGIA E SOCIEDADE DO RIO DE JANEIRO (ITS Rio); CENTRO BRASILEIRO DE RELAÇÕES INTERNACIONAIS (CEBRI). *Um AI Safety and Opportunities Institute para o Brasil: um desenho de características, funções e potenciais desafios*, pp. 9, 30. Rio de Janeiro. Mar. 2025. Available at: https://itsrio.org/en/publicacoes/an-ai-safety-and-opportunities-institute-for-brazil-a-design-of-features-functions-and-potential-challenges/. Accessed: 14 May 2026; ARAUJO, Renan; FORT, Kristina; GUEST, Oliver. *Understanding the First Wave of AI Safety Institutes: Characteristics, Functions, and Challenges*, pp. 11-13. Oxford: Institute for AI Policy and Strategy (IAPS). Oct. 2024. Available at: https://doi.org/10.48550/arXiv.2410.09219. Accessed: 22 May 2026.

3 THURNHERR, Lara. *Which Information should the UK and US AISI share with an International Network of AISIs?* London: King's College London, p. 2. Mar. 2025. Available at: https://arxiv.org/abs/2503.04741. Accessed: 27 May 2026.

4 MINISTRY OF SCIENCE, TECHNOLOGY AND INNOVATION (MCTI); CENTRE FOR STRATEGIC STUDIES AND MANAGEMENT (CGEE). *AI for the Good of All: Brazilian Artificial Intelligence Plan*. 2024. Available at: https://pbia.cgee.org.br/. Accessed: 19 May 2026.

5 INSTITUTO DE TECNOLOGIA E SOCIEDADE DO RIO DE JANEIRO (ITS Rio); CENTRO BRASILEIRO DE RELAÇÕES INTERNACIONAIS (CEBRI). *Um AI Safety and Opportunities Institute para o Brasil: um desenho de características, funções e potenciais desafios*, pp. 4, 31. Rio de Janeiro. Mar. 2025. Available at: https://itsrio.org/en/publicacoes/an-ai-safety-and-opportunities-institute-for-brazil-a-design-of-features-functions-and-potential-challenges/. Accessed: 14 May 2026.

developed and deployed in the Brazilian context, and to treat that question with the same analytical seriousness that the international literature currently reserves for frontier risk. Safety and opportunity, in this design, are not competing priorities but complementary registers of the same institutional mission.

The paper proceeds by first examining, in Section 2, the first wave of AISIs, tracing their origins in the Bletchley Park AI Safety Summit process, their subsequent consolidation through successive international summits, and the institutional characteristics and functions that have progressively converged across eight jurisdictions. Section 3 then analyses the mismatch between the developmental conditions presupposed by these institutions and Brazil's current institutional and technological position. Building upon this diagnosis, Section 4 advances the AISO model and outlines its proposed characteristics and functions in light of Brazil's existing institutional infrastructure.

2. AISIS AND THE FIRST WAVE

2.1 What are AISIs

AISIs are a recent and still-emerging institutional category. It is important to note that the first wave of AISIs is the category referent to the institutions that have been established so far, the second wave is still an evolving concept.[6]

The expression "first wave" comes from the policy literature, where it is used both in a strict and in a loose sense. In the strict sense, the term refers to the three bodies announced between November 2023 and February 2024 in the United Kingdom, the United States, and Japan, which share a tightly aligned set of characteristics and

[6] INSTITUTO DE TECNOLOGIA E SOCIEDADE DO RIO DE JANEIRO (ITS Rio); CENTRO BRASILEIRO DE RELAÇÕES INTERNACIONAIS (CEBRI). *Um AI Safety and Opportunities Institute para o Brasil: um desenho de características, funções e potenciais desafios*. Rio de Janeiro: ITS Rio; CEBRI, Mar. 2025, p.8. Available at: https://itsrio.org/en/publicacoes/an-ai-safety-and-opportunities-institute-for-brazil-a-design-of-features-functions-and-potential-challenges/. Accessed: 14 May 2026.

which were directly inspired by the Bletchley process.[7] In the looser sense, it covers the broader group of jurisdictions that established AISIs or functionally equivalent bodies during the same period, including Canada, Singapore, France, South Korea, and the European Union.[8]

From this foundation, scholars have begun projecting what the next wave will look like. This section situates itself within that framework, examining where the first-wave AISI model falls short for Brazil – and for the Global South more broadly – and proposing an institutional design that addresses AI safety challenges without foreclosing the opportunities the technology presents.

As outlined by the *Institute for AI Policy and Strategy* , the first generation of AISIS [9] is generally defined by three fundamental characteristics: (i) \technical institutions, (ii) inside State infrastructure, and (iii) that operate focused on the safety of advanced AI systems. Furthermore, AISIs were originally designed as bodies without binding regulatory powers, whose authority derives from scientific credibility, convening capacity, and participation in a reciprocal international network rather than from the ability to issue enforceable rules.[10] An AISI is thus simultaneously governmental and

---

[7] ARAUJO, Renan; FORT, Kristina; GUEST, Oliver. *Understanding the First Wave of AI Safety Institutes: Characteristics, Functions, and Challenges*. Oxford: Institute for AI Policy and Strategy (IAPS), pp. 5-8. 2024. Available at: https://doi.org/10.48550/arXiv.2410.09219. Accessed: 22 May 2026.

[8] INSTITUTO DE TECNOLOGIA E SOCIEDADE DO RIO DE JANEIRO (ITS Rio); CENTRO BRASILEIRO DE RELAÇÕES INTERNACIONAIS (CEBRI). *Um AI Safety and Opportunities Institute para o Brasil: um desenho de características, funções e potenciais desafios*. Rio de Janeiro: ITS Rio; CEBRI, Mar. 2025, p. 9, 20-25. Available at: https://itsrio.org/en/publicacoes/an-ai-safety-and-opportunities-institute-for-brazil-a-design-of-features-functions-and-potential-challenges/. Accessed: 14 May 2026; ALLEN, Gregory C.; ADAMSON, Georgia. *The AI Safety Institute International Network: Next Steps and Recommendations*. Washington, D.C.: Center for Strategic and International Studies (CSIS), 2024. Available at: https://www.csis.org/analysis/ai-safety-institute-international-network-next-steps-and-recommendations. Accessed: 19 May 2026.

[9] ARAUJO, Renan; FORT, Kristina; GUEST, Oliver. *Understanding the First Wave of AI Safety Institutes: Characteristics, Functions, and Challenges*. Oxford: Institute for AI Policy and Strategy (IAPS). 2024. Available at: https://doi.org/10.48550/arXiv.2410.09219. Accessed: 22 May 2026.

[10] ARAUJO, Renan; FORT, Kristina; GUEST, Oliver. *Understanding the First Wave of AI Safety Institutes: Characteristics, Functions, and Challenges*. Oxford: Institute for AI Policy and Strategy

technical: embedded in the State apparatus, yet structurally distinct from the regulatory agencies that more commonly populate emerging technology fields.

So far, the empirical core of AISI activity has been safety evaluations of frontier models: techniques that test advanced AI systems against tasks relevant to risks such as cyber, chemical, and biological misuse, and that yield documented findings on system behaviour and capabilities.[11] Around this empirical core, three further functions have developed in convergent fashion across the bodies founded between late 2023 and 2024: (i) research aimed at advancing the underlying science of AI safety; (ii) the development of standards capable of shaping market practices and protocols within and beyond government; and (iii) cooperation, both domestically – across academia, industry and civil society – and internationally, through the network of AISIs formalised in San Francisco in November 2024.[12] This triad reflects an understanding that technical evidence, voluntarily produced and openly circulated, can influence the development trajectory of frontier AI without recourse to the regulatory mechanisms typically associated with high-risk technologies.

2.2 The AISIs' Origin

Between 2023 and 2026, a sequence of international AI summits shaped both the diplomatic vocabulary and the institutional architecture of the emerging AISI ecosystem. While Bletchley established the frontier-safety framing that defined the first wave, subsequent summits broadened the agenda to encompass industrial policy, deployment, and Global South priorities.

(IAPS), p. 1. 2024. Available at: https://doi.org/10.48550/arXiv.2410.09219. Accessed: 22 May 2026.

11 Ibidem, pp. 1, 17–19.

12 ARAUJO, Renan; FORT, Kristina; GUEST, Oliver. *Understanding the First Wave of AI Safety Institutes: Characteristics, Functions, and Challenges*. Oxford: Institute for AI Policy and Strategy, pp. 19–30; INTERNATIONAL NETWORK OF AI SAFETY INSTITUTES. *Mission Statement*. San Francisco. 21 Nov. 2024. Available at: https://www.nist.gov/system/files/documents/2024/11/20/Mission%20Statement%20-%20International%20Network%20of%20AISIs.pdf. Accessed: 17 May 2026.

2.2.1 UK AI Safety Summit

The first AI Safety Summit was held at Bletchley Park, in November 2023, convened by the UK and attended by twenty-eight countries and the EU.[13] The choice of venue, once the home of Alan Turing's wartime codebreaking work, was symbolically aligned with the summit's ambition: to inaugurate a new diplomatic conversation about advanced AI at the level of foundational institutional design.[14]

This summits formal objectives centred on frontier AI, encompassing highly capable general-purpose models, including foundation models, that could match or exceed the capabilities of the most advanced systems then in existence and on the catastrophic risks associated with them, particularly in cybersecurity, biotechnology, large-scale disinformation, and alignment failures.[15]

The principal output of the summit was the Bletchley Declaration, endorsed by a broad coalition that included the USA, the UK, the EU, Brazil, India, China, Indonesia, etc.[16] The Declaration committed signatories to identifying AI safety risks of

---

[13] UK GOVERNMENT. *About the AI Safety Summit 2023*. GOV.UK. 2023. Available at: https://www.gov.uk/government/topical-events/ai-safety-summit-2023/about. Accessed: 13 May 2026; UK GOVERNMENT. *Countries Agree to Safe and Responsible Development of Frontier AI in Landmark Bletchley Declaration*. GOV.UK. 1 Nov. 2023. Available at: https://www.gov.uk/government/news/countries-agree-to-safe-and-responsible-development-of-frontier-ai-in-landmark-bletchley-declaration. Accessed: 15 May 2026.

[14] CUÉLLAR, Mariano-Florentino. *The UK AI Safety Summit Opened a New Chapter in AI Diplomacy*. Washington, D.C.: Carnegie Endowment for International Peace. Nov. 2023. Available at: https://carnegieendowment.org/posts/2023/11/the-uk-ai-safety-summit-opened-a-new-chapter-in-ai-diplomacy. Accessed: 20 May 2026.

[15] UK GOVERNMENT. *The Bletchley Declaration by Countries Attending the AI Safety Summit, 1-2 November 2023*. GOV.UK. 1-2 Nov. 2023. Available at: https://www.gov.uk/government/publications/ai-safety-summit-2023-the-bletchley-declaration/the-bletchley-declaration-by-countries-attending-the-ai-safety-summit-1-2-november-2023. Accessed: 16 May 2026.

[16] UK GOVERNMENT. *The Bletchley Declaration by Countries Attending the AI Safety Summit, 1-2 November 2023*. GOV.UK. 1-2 Nov. 2023. Available at: https://www.gov.uk/government/publications/ai-safety-summit-2023-the-bletchley-declaration/the-bletchley-declaration-by-countries-attending-the-ai-safety-summit-1-2-november-2023. Accessed: 16 May 2026; INSTITUTO DE TECNOLOGIA E SOCIEDADE DO RIO DE JANEIRO (ITS Rio); CENTRO BRASILEIRO

shared concern, building a shared evidence-based understanding of those risks, and supporting an internationally inclusive network of scientific research on frontier AI safety.[17] This last commitment, understated in the text itself, proved institutionally consequential: it laid the groundwork both for what would become the International AI Safety Report, led by Yoshua Bengio, and for the network of national AISIs that would emerge in the following months.[18] At Bletchley itself, the UK and the USA announced the first two national AI Safety Institutes – technical, advisory, embedded within government, and explicitly without regulatory powers.[19]

It is from this founding moment that the so-called "Bletchley effect" derives – the cascade of national AISIs established or announced in jurisdictions inspired by the British precedent: Japan, Singapore, Canada, France, South Korea, and, in function-

---

DE RELAÇÕES INTERNACIONAIS (CEBRI). *Um AI Safety and Opportunities Institute para o Brasil: um desenho de características, funções e potenciais desafios*. Rio de Janeiro: ITS Rio; CEBRI, Mar. 2025, p. 10, 20-25. Available at: https://itsrio.org/en/publicacoes/an-ai-safety-and-opportunities-institute-for-brazil-a-design-of-features-functions-and-potential-challenges/. Accessed: 14 May 2026.; ALLEN, Gregory C.; ADAMSON, Georgia. *The AI Safety Institute International Network: Next Steps and Recommendations*. Washington, D.C.: Center for Strategic and International Studies (CSIS), p. 10, 2024. Available at: https://www.csis.org/analysis/ai-safety-institute-international-network-next-steps-and-recommendations. Accessed: 19 May 2026.

17 UK GOVERNMENT. *The Bletchley Declaration by Countries Attending the AI Safety Summit, 1-2 November 2023*. GOV.UK. 1-2 Nov. 2023. Available at: https://www.gov.uk/government/publications/ai-safety-summit-2023-the-bletchley-declaration/the-bletchley-declaration-by-countries-attending-the-ai-safety-summit-1-2-november-2023 . Accessed: 16 May 2026.

18 INTERNATIONAL AI SAFETY REPORT. *International AI Safety Report 2026*. 2026. Available at: https://internationalaisafetyreport.org/publication/international-ai-safety-report-2026. Accessed: 24 May 2026; LEONE, Daniela et al. *Evaluating the Potential Functions of an International Institution for AI Safety: An Institutional Analysis of the IAEA and IPCC, in the Context of Recent Trends in AI Safety*. London: The Alan Turing Institute. arXiv:2409.10536v1. Sept. 2024. Available at: https://doi.org/10.48550/arXiv.2409.10536. Accessed: 26 May 2026.

19 ARAUJO, Renan; FORT, Kristina; GUEST, Oliver. *Understanding the First Wave of AI Safety Institutes: Characteristics, Functions, and Challenges*. Oxford: Institute for AI Policy and Strategy (IAPS), p. 7. 2024. Available at: https://doi.org/10.48550/arXiv.2410.09219. Accessed: 22 May 2026; UK GOVERNMENT. *Introducing the AI Safety Institute*. Nov. 2023. Available at: https://assets.publishing.service.gov.uk/media/65438d159e05fd0014be7bd9/introducing-ai-safety-institute-web-accessible.pdf. Accessed: 12 May 2026.

ally adjacent form, the EU.[20]

2.2.2 AI Seoul Summit

The second meeting was held May 2024, co-hosted by the Republic of Korea and the UK. The Seoul Summit was framed from the outset as a continuation of Bletchley, but it introduced a widening of the conceptual frame. The summit's three priorities were safety, innovation, and inclusivity. This indicated that the international conversation on AI governance could no longer be organised exclusively around the catastrophic-risk vocabulary that had shaped the British meeting.[21]

The document architecture itself reflected this shift: rather than a single declaration, the summit produced three distinct outputs, each addressing a different dimension of the emerging consensus.[22]

The first of these, the Seoul Declaration for Safe, Innovative and Inclusive AI,

---

[20] INSTITUTO DE TECNOLOGIA E SOCIEDADE DO RIO DE JANEIRO (ITS Rio); CENTRO BRASILEIRO DE RELAÇÕES INTERNACIONAIS (CEBRI). *Um AI Safety and Opportunities Institute para o Brasil: um desenho de características, funções e potenciais desafios*. Rio de Janeiro: ITS Rio; CEBRI, Mar. 2025, p. 10, 20-25. Available at: https://itsrio.org/en/publicacoes/an-ai-safety-and-opportunities-institute-for-brazil-a-design-of-features-functions-and-potential-challenges/. Accessed: 14 May 2026; HERN, Alex. Trying to Tame AI: Seoul Summit Flags Hurdles to Regulation. *The Guardian*. 27 May 2024. Available at: https://www.theguardian.com/technology/article/2024/may/27/trying-to-tame-ai-seoul-summit-flags-hurdles-to-regulation. Accessed: 23 May 2026.

[21] UK GOVERNMENT. *About the AI Seoul Summit 2024*. GOV.UK. 2024. Available at: https://www.gov.uk/government/topical-events/ai-seoul-summit-2024/about. Accessed: 11 May 2026; UK GOVERNMENT. *Seoul Declaration for Safe, Innovative and Inclusive AI*. AI Seoul Summit. 21 May 2024. Available at: https://www.gov.uk/government/publications/seoul-declaration-for-safe-innovative-and-inclusive-ai-ai-seoul-summit-2024. Accessed: 28 May 2026.

[22] INSTITUTO DE TECNOLOGIA E SOCIEDADE DO RIO DE JANEIRO (ITS Rio); CENTRO BRASILEIRO DE RELAÇÕES INTERNACIONAIS (CEBRI). *Um AI Safety and Opportunities Institute para o Brasil: um desenho de características, funções e potenciais desafios*. Rio de Janeiro: ITS Rio; CEBRI, Mar. 2025, p. 11, 20-25. Available at: https://itsrio.org/en/publicacoes/an-ai-safety-and-opportunities-institute-for-brazil-a-design-of-features-functions-and-potential-challenges/. Accessed: 20 May 2026.

recognized that "AI safety, innovation, and inclusivity are interrelated goals".[23] The second, The Seoul Ministerial Statement articulated a more detailed framework for international cooperation on AI governance, including commitments to support the creation or expansion of AISIs, research programmes, and other supervisory bodies.[24]

The third document, the Seoul Statement of Intent toward International Cooperation on AI Safety Science established the foundations for what would later become the International Network of AI Safety Institutes.[25] This document provided the formal basis for the multilateral network of AISIs that followed.

2.2.3 AI Action Summit (Paris)

The Paris meeting was held in February 2025, hosted by France and co-chaired by India. The British and South Korean summits had been called AI Safety Summit and AI Seoul Summit; the French gathering was deliberately renamed the AI Action Summit, signalling a programmatic intent to move from the sole purpose of risk identification towards subjects such as implementation, competitiveness, and innovation.[26]

---

[23] UK GOVERNMENT. *Seoul Declaration for Safe, Innovative and Inclusive AI*. AI Seoul Summit. 21 May 2024. Available at: https://www.gov.uk/government/publications/seoul-declaration-for-safe-innovative-and-inclusive-ai-ai-seoul-summit-2024. Accessed: 28 May 2026.

[24] Ibidem.

[25] UK GOVERNMENT. *Seoul Statement of Intent Toward International Cooperation on AI Safety Science*. AI Seoul Summit 2024. GOV.UK. 2024. Available at: https://www.gov.uk/government/publications/seoul-declaration-for-safe-innovative-and-inclusive-ai-ai-seoul-summit-2024/seoul-statement-of-intent-toward-international-cooperation-on-ai-safety-science-ai-seoul-summit-2024-annex; ARAUJO, Renan; FORT, Kristina; GUEST, Oliver. *Understanding the First Wave of AI Safety Institutes: Characteristics, Functions, and Challenges*. Oxford: Institute for AI Policy and Strategy (IAPS), p. 32. 2024. Available at: https://doi.org/10.48550/arXiv.2410.09219. Accessed: 22 May 2026; INTERNATIONAL NETWORK OF AI SAFETY INSTITUTES. *Mission Statement*. San Francisco. 21 Nov. 2024. Available at: https://www.nist.gov/system/files/documents/2024/11/20/Mission%20Statement%20-%20International%20Network%20of%20AISIs.pdf. Accessed: 17 May 2026.

[26] ÉLYSÉE PALACE. *AI Action Summit*. Paris. Feb. 2025. Available at: https://www.elysee.fr/en/sommet-pour-l-action-sur-l-ia. Accessed: 18 May 2026; INSTITUTO DE

During the meeting, the French President framed AI as a geopolitical and economic asset rather than a regulatory challenge, emphasising French nuclear infrastructure as a competitive advantage for hosting sustainable AI investments. The U.S. Vice President, attending in the early weeks of the new Trump administration, was more explicit still: "I'm not here this morning to talk about AI safety, which was the title of the conference a couple of years ago. I'm here to talk about AI opportunity". [27]

Paris contained the Seoul seeds of a different institutional vocabulary. Discussions on sustainability, energy infrastructure, and the integration of AI with development priorities pointed toward a broader institutional agenda that the first-wave AISI model had never been designed to absorb. Countries with renewable energy potential, Brazil among them, were explicitly encouraged to position themselves as hosts of sustainable AI investment.[28]

2.2.4 India AI Impact Summit

The fourth gathering took place in February 2026 in New Delhi,[29] and India

TECNOLOGIA E SOCIEDADE DO RIO DE JANEIRO (ITS Rio); CENTRO BRASILEIRO DE RELAÇÕES INTERNACIONAIS (CEBRI). *Um AI Safety and Opportunities Institute para o Brasil: um desenho de características, funções e potenciais desafios*. Rio de Janeiro: ITS Rio; CEBRI, Mar. 2025, p. 12, 20-25. Available at: https://itsrio.org/en/publicacoes/an-ai-safety-and-opportunities-institute-for-brazil-a-design-of-features-functions-and-potential-challenges/. Accessed: 21 May 2026.

27 PERRIGO, Billy. U.S. Pushes for Less AI Regulation at Paris Summit. *TIME*. Feb. 2025. Available at: https://time.com/7221384/ai-regulation-takes-bac; POUGET, Hadrien. *France's AI Summit Is a Chance to Reshape Global Narratives on AI*. Washington, D.C.: Carnegie Endowment for International Peace. July 2024. Available at: https://carnegieendowment.org/posts/2024/07/france-ai-summit-reshape-global-narrative. Accessed: 21 May 2026.

28 INSTITUTO DE TECNOLOGIA E SOCIEDADE DO RIO DE JANEIRO (ITS Rio); CENTRO BRASILEIRO DE RELAÇÕES INTERNACIONAIS (CEBRI). *Um AI Safety and Opportunities Institute para o Brasil: um desenho de características, funções e potenciais desafios*. Rio de Janeiro: ITS Rio; CEBRI, Mar. 2025, p. 12, 20-25. Available at: https://itsrio.org/en/publicacoes/an-ai-safety-and-opportunities-institute-for-brazil-a-design-of-features-functions-and-potential-challenges/. Accessed: 22 May 2026.

29 KRISHNAN, Murali. India's AI Governance Push Takes Center Stage at Summit. *Deutsche*

became the first developing country to host a global AI summit. The hosting was framed by the Indian government as a deliberate assertion of voice on behalf of the Global South.[30]

The summit's name continued the trajectory of relabelling. "Safety" had given way to "action" in Paris, and "action" gave way to "impact". In his inaugural address, the Indian Prime Minister framed the meeting around the question of "how to make AI from machine-centric to human-centric; how to make it sensitive and responsive".[31]

It was set by an agenda reflecting the priorities of states whose principal stake in AI governance is not only the prevention of frontier misuse but the responsible deployment of existing systems at scale. Conversations at the summit foregrounded affordable deployment, indigenous large-language models, and AI applications in healthcare and education, in dialogue with the broader concern that the AISI model as originally conceived was too narrow to address the priorities of countries whose challenges lay closer to agriculture, health infrastructure, and digital inclusion than to alignment failures in frontier systems.[32]

For a section concerned with Brazilian possibilities, the India summit carries particular relevance. It demonstrates that the second-wave institutional design hinted at in Paris is no longer hypothetical. The presence of Brazilian President at the head table is itself a fact worth registering: Brazil arrives at the threshold of any future AISI conversation already embedded in the highest-level diplomatic forum on AI governance, and on the side of the conversation that has now visibly moved beyond

---

Welle. 20 Feb. 2026. Available at: https://www.dw.com/en/indias-ai-governance-push-takes-center-stage-at-summit/a-76058826. Accessed: 27 May 2026;

30 Ibidem.

31 Ibidem.

32 CHAUDHURI, Rudra. *Disrupting AI Safety Institutes: The India Way*. Washington, D.C.: Carnegie Endowment for International Peace. Sept. 2024. Available at: https://carnegieendowment.org/posts/2024/09/disrupting-ai-safety-institutes-the-india-way. Accessed: 24 May 2026; ARAUJO, Renan; FORT, Kristina; GUEST, Oliver. *Understanding the First Wave of AI Safety Institutes: Characteristics, Functions, and Challenges*. Oxford: Institute for AI Policy and Strategy (IAPS). 2024. Available at: https://doi.org/10.48550/arXiv.2410.09219. Accessed: 22 May 2026; KRISHNAN, Murali. India's AI Governance Push Takes Center Stage at Summit. *Deutsche Welle*. 20 Feb. 2026. Available at: https://www.dw.com/en/indias-ai-governance-push-takes-center-stage-at-summit/a-76058826. Accessed: 27 May 2026.

its founding Anglo-American framing.

2.3 AISIs - First wave

The eight jurisdictions to be examined in this section share a common emphasis on technical evaluation of advanced AI systems, but they differ in certain aspects in their institutional design. In order to suggest a Brazilian model, it is important to note what has been established so far and, from it, build what will best apply to the country's reality.

2.3.1 United Kingdom

The UK established its institute in November 2023, announced at the Bletchley Summit.[33] It emerged from the renaming of the Frontier AI Taskforce, originally launched in April 2023, and was placed within the Department for Science, Innovation and Technology, with Ian Hogarth as Chair.[34]

The mission, as articulated at launch, was to "minimise surprise to the UK and humanity from rapid and unexpected advances in AI" by developing the sociotechnical infrastructure needed to understand the risks of advanced AI systems and to enable their governance.[35] In practice this has meant three principal lines of work:

---

[33] UK GOVERNMENT. *Introducing the AI Safety Institute*. Nov. 2023. Available at: https://assets.publishing.service.gov.uk/media/65438d159e05fd0014be7bd9/introducing-ai-safety-institute-web-accessible.pdf. Accessed: 12 May 2026.

[34] ARAUJO, Renan; FORT, Kristina; GUEST, Oliver. *Understanding the First Wave of AI Safety Institutes: Characteristics, Functions, and Challenges*. Oxford: Institute for AI Policy and Strategy (IAPS), pp. 6-8. 2024. Available at: https://doi.org/10.48550/arXiv.2410.09219. Accessed: 22 May 2026; INSTITUTO DE TECNOLOGIA E SOCIEDADE DO RIO DE JANEIRO (ITS Rio); CENTRO BRASILEIRO DE RELAÇÕES INTERNACIONAIS (CEBRI). *Um AI Safety and Opportunities Institute para o Brasil: um desenho de características, funções e potenciais desafios*. Rio de Janeiro: ITS Rio; CEBRI, Mar. 2025, pp. 16–17. Available at: https://itsrio.org/en/publicacoes/an-ai-safety-and-opportunities-institute-for-brazil-a-design-of-features-functions-and-potential-challenges/. Accessed: 14 May 2026; UK AI Security Institute, www.aisi.gov.uk/about.

[35] UK GOVERNMENT. *Introducing the AI Safety Institute*. Nov. 2023. Available at: https://assets.publishing.service.gov.uk/media/65438d159e05fd0014be7bd9/introducing-ai-safety-insti-

evaluations of frontier models, foundational research on AI safety, and international coordination, including through the international network of AISIs and the State of the Science Report process.[36]

2.3.2 United States of America

The USA also announced its institute during the Bletchley Summit, in November 2023, and formally established it within the National Institute of Standards and Technology (NIST), an agency of the Department of Commerce.[37] The institutional placement was the NIST, the federal standards-setting body.[38]

The mission of the USA AISI is centred on three lines of work: (i) development of empirically grounded tests, benchmarks and evaluations for advanced models, systems and agents; (ii) performance of pre-deployment test, evaluation, validation and

---

tute-web-accessible.pdf. Accessed: 12 May 2026.

[36] UK AI Security Institute, www.aisi.gov.uk/about; INTERNATIONAL AI SAFETY REPORT. *International AI Safety Report 2026*, pp. 6-8, 2026. Available at: https://internationalaisafetyreport.org/publication/international-ai-safety-report-2026. Accessed: 24 May 2026;

[37] CENTER FOR STRATEGIC AND INTERNATIONAL STUDIES (CSIS). *The AI Safety Institute Network*. Washington, D.C.: CSIS. pp. 8, 12. Oct. 2024. Available at: https://www.csis.org/analysis/ai-safety-institute-international-network-next-steps-and-recommendations. Accessed: 19 May 2026; INSTITUTO DE TECNOLOGIA E SOCIEDADE DO RIO DE JANEIRO (ITS Rio); CENTRO BRASILEIRO DE RELAÇÕES INTERNACIONAIS (CEBRI). *Um AI Safety and Opportunities Institute para o Brasil: um desenho de características, funções e potenciais desafios*. Rio de Janeiro: ITS Rio; CEBRI, Mar. 2025, pp. 16–17. Available at: https://itsrio.org/en/publicacoes/an-ai-safety-and-opportunities-institute-for-brazil-a-design-of-features-functions-and-potential-challenges/. Accessed: 14 May 2026.

[38] ARAUJO, Renan; FORT, Kristina; GUEST, Oliver. *Understanding the First Wave of AI Safety Institutes: Characteristics, Functions, and Challenges*. Oxford: Institute for AI Policy and Strategy (IAPS), p. 12. 2024. Available at: https://doi.org/10.48550/arXiv.2410.09219. Accessed: 22 May 2026; INSTITUTO DE TECNOLOGIA E SOCIEDADE DO RIO DE JANEIRO (ITS Rio); CENTRO BRASILEIRO DE RELAÇÕES INTERNACIONAIS (CEBRI). *Um AI Safety and Opportunities Institute para o Brasil: um desenho de características, funções e potenciais desafios*. Rio de Janeiro: ITS Rio; CEBRI, Mar. 2025, pp. 18–19, 30. Available at: https://itsrio.org/en/publicacoes/an-ai-safety-and-opportunities-institute-for-brazil-a-design-of-features-functions-and-potential-challenges/. Accessed: 14 May 2026

verification of frontier models; and (iii) active participation in domestic and international standard-setting processes.[39] The institute has signed cooperation agreements with Anthropic and OpenAI that grant it pre-deployment access to leading models. Its work feeds directly into the broader NIST AI Risk Management Framework, a document that has influenced the AI governance vocabulary well beyond the United States, including through harmonisation efforts undertaken by the Japan AISI.[40]

2.3.3 Singapore

The Singaporean institute was formally designated in May 2024, when the Infocomm Media Development Authority announced that the Digital Trust Centre (DTC), hosted at Nanyang Technological University, would serve as the country's AI Safety Institute.[41] The Singaporean AISI therefore did not emerge as a new agency.

---

[39] ALLEN, Gregory C.; ADAMSON, Georgia. *The AI Safety Institute International Network: Next Steps and Recommendations*. Washington, D.C.: Center for Strategic and International Studies (CSIS), pp. 5-12, 2024. Available at: https://www.csis.org/analysis/ai-safety-institute-international-network-next-steps-and-recommendations. Accessed: 19 May 2026; INSTITUTO DE TECNOLOGIA E SOCIEDADE DO RIO DE JANEIRO (ITS Rio); CENTRO BRASILEIRO DE RELAÇÕES INTERNACIONAIS (CEBRI). *Um AI Safety and Opportunities Institute para o Brasil: um desenho de características, funções e potenciais desafios*. Rio de Janeiro: ITS Rio; CEBRI, Mar. 2025, p. 18. Available at: https://itsrio.org/en/publicacoes/an-ai-safety-and-opportunities-institute-for-brazil-a-design-of-features-functions-and-potential-challenges/. Accessed: 27 May 2026.

[40] ARAUJO, Renan; FORT, Kristina; GUEST, Oliver. *Understanding the First Wave of AI Safety Institutes: Characteristics, Functions, and Challenges*. Oxford: Institute for AI Policy and Strategy (IAPS), pp. 22, 26 e 28. 2024. Available at: https://doi.org/10.48550/arXiv.2410.09219. Accessed: 22 May 2026; INSTITUTO DE TECNOLOGIA E SOCIEDADE DO RIO DE JANEIRO (ITS Rio); CENTRO BRASILEIRO DE RELAÇÕES INTERNACIONAIS (CEBRI). *Um AI Safety and Opportunities Institute para o Brasil: um desenho de características, funções e potenciais desafios*. Rio de Janeiro: ITS Rio; CEBRI, Mar. 2025, p. 29. Available at: https://itsrio.org/en/publicacoes/an-ai-safety-and-opportunities-institute-for-brazil-a-design-of-features-functions-and-potential-challenges/. Accessed: 14 May 2026.

[41] ARAUJO, Renan; FORT, Kristina; GUEST, Oliver. *Understanding the First Wave of AI Safety Institutes: Characteristics, Functions, and Challenges*. Oxford: Institute for AI Policy and Strategy (IAPS), pp. 8-9. 2024. Available at: https://doi.org/10.48550/arXiv.2410.09219. Accessed: 22 May 2026; DIGITAL TRUST CENTRE (DTC). *About Us*. Singapore: Nanyang Technological Universi-

It was layered onto an existing research centre whose institutional placement, funding, and academic anchoring were already in place.

The mandate of the Singaporean AISI is centred on technical testing and evaluation of AI systems, the development of frameworks for safe design, deployment, and operational integration of AI, and the establishment of content assurance protocols suited to the country's linguistic and cultural context.[42] The DTC also engages in international cooperation, including a 2024 partnership with the EU focused on joint research and standardised testing protocols, and it represents Singapore in the International Network of AISIs.[43]

2.3.4 Canada

The Canadian Artificial Intelligence Safety Institute (CAISI) was launched in November 2024, housed within the federal Department of Innovation, Science and Economic Development of Canada.[44] Rather than operate as a single centralised

ty (NTU). 2024. Available at: https://www.ntu.edu.sg/dtc/about-us. Accessed: 16 May 2026.

42 INSTITUTO DE TECNOLOGIA E SOCIEDADE DO RIO DE JANEIRO (ITS Rio); CENTRO BRASILEIRO DE RELAÇÕES INTERNACIONAIS (CEBRI). *Um AI Safety and Opportunities Institute para o Brasil: um desenho de características, funções e potenciais desafios*. Rio de Janeiro: ITS Rio; CEBRI, Mar. 2025, p. 20. Available at: https://itsrio.org/en/publicacoes/an-ai-safety-and-opportunities-institute-for-brazil-a-design-of-features-functions-and-potential-challenges/. Accessed: 27 May 2026.

43 INSTITUTO DE TECNOLOGIA E SOCIEDADE DO RIO DE JANEIRO (ITS Rio); CENTRO BRASILEIRO DE RELAÇÕES INTERNACIONAIS (CEBRI). *Um AI Safety and Opportunities Institute para o Brasil: um desenho de características, funções e potenciais desafios*. Rio de Janeiro: ITS Rio; CEBRI, Mar. 2025, p. 20. Available at: https://itsrio.org/en/publicacoes/an-ai-safety-and-opportunities-institute-for-brazil-a-design-of-features-functions-and-potential-challenges/. Accessed: 27 May 2026; ARAUJO, Renan; FORT, Kristina; GUEST, Oliver. *Understanding the First Wave of AI Safety Institutes: Characteristics, Functions, and Challenges*. Oxford: Institute for AI Policy and Strategy (IAPS), p.9. 2024. Available at: https://doi.org/10.48550/arXiv.2410.09219. Accessed: 22 May 2026.

44 INNOVATION, SCIENCE AND ECONOMIC DEVELOPMENT CANADA. *Canada Launches Canadian Artificial Intelligence Safety Institute*. 12 Nov. 2024. Available at: https://www.canada.ca/en/innovation-science-economic-development/news/2024/11/canada-launches-canadian-artificial-intelligence-safety-institute.html. Accessed: 20 May 2026; INSTITUTO DE

body, CAISI was structured as a two-stream organisation: (i) the applied and investigator-led research stream, delivered through a contribution agreement with the Canadian Institute for Advanced Research, which coordinates its work with the country's three National AI Institutes (Mila in Montréal, the Vector Institute in Toronto, and Amii in Edmonton); and (ii) government-directed projects, responsible for addressing national priorities such as cybersecurity, is implemented by the National Research Council of Canada.[45]

Overall, CAISI's mission is articulated around the assessment of AI risks, the testing of advanced AI systems, the development of guidance for industry and government, and active engagement with the International Network of AISIs.[46] Its research priorities include detection of AI-generated content, the evaluation of advanced AI models, safe adoption of AI in high-risk applications, and privacy preservation in AI systems.

2.3.5 Japan

The Japanese AI Safety Institute was launched in February 2024, established

TECNOLOGIA E SOCIEDADE DO RIO DE JANEIRO (ITS Rio); CENTRO BRASILEIRO DE RELAÇÕES INTERNACIONAIS (CEBRI). *Um AI Safety and Opportunities Institute para o Brasil: um desenho de características, funções e potenciais desafios*. Rio de Janeiro: ITS Rio; CEBRI, Mar. 2025, p. 21. Available at: https://itsrio.org/en/publicacoes/an-ai-safety-and-opportunities-institute-for-brazil-a-design-of-features-functions-and-potential-challenges/. Accessed: 27 May 2026.

45 INNOVATION, SCIENCE AND ECONOMIC DEVELOPMENT CANADA. *Canada Launches Canadian Artificial Intelligence Safety Institute*. 12 Nov. 2024. Available at: https://www.canada.ca/en/innovation-science-economic-development/news/2024/11/canada-launches-canadian-artificial-intelligence-safety-institute.html. Accessed: 20 May 2026; CANADIAN INSTITUTE FOR ADVANCED RESEARCH (CIFAR). *CIFAR to Lead an AI Safety Research Program in Support of the Canadian AI Safety Institute*. 12 Nov. 2024. Available at: https://cifar.ca/cifar-news/2024/11/12/government-of-canada-announces-canadian-ai-safety-institute/. Accessed: 22 May 2026.

46 INNOVATION, SCIENCE AND ECONOMIC DEVELOPMENT CANADA. *Canada Launches Canadian Artificial Intelligence Safety Institute*. 12 Nov. 2024. Available at: https://www.canada.ca/en/innovation-science-economic-development/news/2024/11/canada-launches-canadian-artificial-intelligence-safety-institute.html. Accessed: 20 May 2026

within the Information-technology Promotion Agency, an organisation under the Ministry of Economy, Trade and Industry, with Akiko Murakami as its inaugural director.[47] It operates in collaboration with a particularly wide range of relevant ministries, including the Cabinet Secretariat, the National Police Agency, the Digital Agency, the Ministry of Foreign Affairs, the Ministry of Defense, and several specialised research institutes such as the National Institute of Advanced Industrial Science and Technology and the National Institute of Informatics.[48]

The Japanese AISI's purpose is centred on three principal responsibilities: research and examination of criteria and standards for AI safety evaluation, examination of evaluation methods, and international collaboration with peer institutes, particularly those in the USA and the UK.[49] Rather than concentrate evaluative capacity in a single resource-intensive body, Japan distributed responsibilities across a network of pre-existing research institutes and ministries, with the AISI functioning as the coordinating node.[50]

2.3.6 France

In France the *Institut National pour l'Évaluation et la Sécurité de l'Intelligence*

---

47 MINISTRY OF ECONOMY, TRADE AND INDUSTRY (METI). *Launch of AI Safety Institute*. 14 Feb. 2024. Available at: https://www.meti.go.jp/english/press/2024/0214_001.html. Accessed: 18 May 2026; ARAUJO, Renan; FORT, Kristina; GUEST, Oliver. *Understanding the First Wave of AI Safety Institutes: Characteristics, Functions, and Challenges*. Oxford: Institute for AI Policy and Strategy (IAPS), p.8. 2024. Available at: https://doi.org/10.48550/arXiv.2410.09219. Accessed: 22 May 2026.

48 MINISTRY OF ECONOMY, TRADE AND INDUSTRY (METI). *Launch of AI Safety Institute*. 14 Feb. 2024. Available at: https://www.meti.go.jp/english/press/2024/0214_001.html. Accessed: 18 May 202.

49 Ibidem.

50 ARAUJO, Renan; FORT, Kristina; GUEST, Oliver. *Understanding the First Wave of AI Safety Institutes: Characteristics, Functions, and Challenges*. Oxford: Institute for AI Policy and Strategy (IAPS), p. 19. 2024. Available at: https://doi.org/10.48550/arXiv.2410.09219. Accessed: 22 May 2026; MINISTRY OF ECONOMY, TRADE AND INDUSTRY (METI). *Launch of AI Safety Institute*. 14 Feb. 2024. Available at: https://www.meti.go.jp/english/press/2024/0214_001.html. Accessed: 18 May 2026.

Artificielle (INESIA), was announced in January 2025 by the French Ministry of Economy and Finance, shortly before the AI Action Summit in Paris.[51] Rather than create a single new agency, France established INESIA as a partnership between two existing research bodies, the *Laboratoire National de Métrologie et d'Essais* (LNE) and the *Institut National de Recherche en Informatique et en Automatique* (INRIA), with a mandate centred on the development of testing and evaluation methodologies for general-purpose AI models at the national level.[52]

So far, France has deliberately positioned itself away from the catastrophic-risk and *safety* vocabulary and toward a vision of AI governance organised around openness, innovation, and the construction of a European third way between the USA and China.[53] A government-appointed commission report largely dismissed the framing of catastrophic frontier risks and argued that such risks were being used by incumbent US firms to legitimise barriers to entry that would consolidate their market

---

51 MINISTÈRE DE L'ÉCONOMIE ET DES FINANCES. *La France se dote d'un Institut national pour l'évaluation et la sécurité de l'intelligence artificielle (INESIA)*. Jan. 2025. Available at: https://www.economie.gouv.fr/actualites/la-france-se-dote-dun-institut-national-pour-levaluation-et-la-securite-de-lintelligence#. Accessed: 15 May 2026.

52 MINISTÈRE DE L'ÉCONOMIE ET DES FINANCES. *La France se dote d'un Institut national pour l'évaluation et la sécurité de l'intelligence artificielle (INESIA)*. Jan. 2025. Available at: https://www.economie.gouv.fr/actualites/la-france-se-dote-dun-institut-national-pour-levaluation-et-la-securite-de-lintelligence#. Accessed: 15 May 2026; ALLEN, Gregory C.; ADAMSON, Georgia. *The AI Safety Institute International Network: Next Steps and Recommendations*. Washington, D.C.: Center for Strategic and International Studies (CSIS), p. 10, 2024. Available at: https://www.csis.org/analysis/ai-safety-institute-international-network-next-steps-and-recommendations. Accessed: 19 May 2026; INSTITUTO DE TECNOLOGIA E SOCIEDADE DO RIO DE JANEIRO (ITS Rio); CENTRO BRASILEIRO DE RELAÇÕES INTERNACIONAIS (CEBRI). *Um AI Safety and Opportunities Institute para o Brasil: um desenho de características, funções e potenciais desafios*. Rio de Janeiro: ITS Rio; CEBRI, Mar. 2025, p. 23. Available at: https://itsrio.org/en/publicacoes/an-ai-safety-and-opportunities-institute-for-brazil-a-design-of-features-functions-and-potential-challenges/. Accessed: 27 May 2026.

53 POUGET, Hadrien. *France's AI Summit Is a Chance to Reshape Global Narratives on AI*. Washington, D.C.: Carnegie Endowment for International Peace. July 2024. Available at: https://carnegieendowment.org/posts/2024/07/france-ai-summit-reshape-global-narrative. Accessed: 21 May 2026.

dominance,[54] illustrating the country's geopolitical standing.

2.3.7 European Union

The EU AI Office was established within the European Commission in February 2024 and began operations under the Directorate-General for Communications Networks, Content and Technology.[55] It is made up of six functional units, including dedicated units for Regulation and Compliance, AI Safety, AI Innovation and Policy Coordination, AI for Societal Good, and AI in Health and Life Sciences.[56] The body sits at the centre of what the European Commission has described as the foundation of a single European AI governance system.[57]

Under the AI Act, the Commission, supported by the AI Office, is empowered to conduct evaluations of general-purpose AI models, request information and corrective measures from model providers, and apply sanctions for non-compliance.[58] In

---

54 POUGET, Hadrien. *France's AI Summit Is a Chance to Reshape Global Narratives on AI*. Washington, D.C.: Carnegie Endowment for International Peace. July 2024. Available at: https://carnegieendowment.org/posts/2024/07/france-ai-summit-reshape-global-narrative. Accessed: 21 May 2026; INSTITUTO DE TECNOLOGIA E SOCIEDADE DO RIO DE JANEIRO (ITS Rio); CENTRO BRASILEIRO DE RELAÇÕES INTERNACIONAIS (CEBRI). *Um AI Safety and Opportunities Institute para o Brasil: um desenho de características, funções e potenciais desafios*. Rio de Janeiro: ITS Rio; CEBRI, Mar. 2025, p. 23. Available at: https://itsrio.org/en/publicacoes/an-ai-safety-and-opportunities-institute-for-brazil-a-design-of-features-functions-and-potential-challenges/. Accessed: 27 May 2026.

55 EUROPEAN COMMISSION. *European AI Office*. 2026. Available at: https://digital-strategy.ec.europa.eu/en/policies/ai-office. Accessed: 13 May 2026; ALLEN, Gregory C.; ADAMSON, Georgia. *The AI Safety Institute International Network: Next Steps and Recommendations*. Washington, D.C.: Center for Strategic and International Studies (CSIS), p. 11. 2024. Available at: https://www.csis.org/analysis/ai-safety-institute-international-network-next-steps-and-recommendations. Accessed: 19 May 2026.

56 EUROPEAN COMMISSION. *European AI Office*. 2026. Available at: https://digital-strategy.ec.europa.eu/en/policies/ai-office. Accessed: 13 May 2026.

57 Ibidem.

58 EUROPEAN COMMISSION. *European AI Office: Shaping Europe's Digital Future*. Brussels. 2024. Available at: https://digital-strategy.ec.europa.eu/en/policies/ai-office. Accessed: 21 May 2026; ARAUJO, Renan; FORT, Kristina; GUEST, Oliver. *Understanding the First Wave of AI*

the analytical terminology developed earlier in this chapter, this is the only first-wave-contemporaneous body whose mandate explicitly fuses the technical evaluative function with binding regulatory authority.

2.3.8 South Korea

The Korean AI Safety Institute was launched in November 2024 at the Pangyo Global R&D Center, just south of Seoul, under the Ministry of Science and ICT and hosted within the Electronics and Telecommunications Research Institute (ETRI).[59]

At its launch ceremony, a Memorandum of Understanding was signed establishing the Korea AI Safety Consortium, a coalition of twenty-four organisations from industry, academia, and research institutes.[60] The industry participants include Naver, KT, Kakao, LG AI Research, SKT, and Samsung Electronics; the academic participants include Seoul National University, KAIST, and Korea University, among others.[61] The Consortium's stated purposes include joint research on AI safety frameworks, alignment with international safety norms, and technological collaboration. This consortium-based design produces a body that is, by structural intent, deeply embedded in the Korean AI industry rather than positioned at arm's

Safety Institutes: Characteristics, Functions, and Challenges. Oxford: Institute for AI Policy and Strategy (IAPS), p. 28. 2024. Available at: https://doi.org/10.48550/arXiv.2410.09219. Accessed: 22 May 2026.

59 KOREA TIMES. *Korea's AI Safety Institute Officially Launches*. 27 Nov. 2024. Available at: https://www.koreatimes.co.kr/business/tech-science/20241127/koreas-ai-safety-institute-officially-launches. Accessed: 23 May 2026.

60 NATIONAL RESEARCH COUNCIL OF SCIENCE AND TECHNOLOGY. A*I Safety Institute Launched as Korea's AI Research Hub*. Newswise. 27 Nov. 2024. Available at:https://www.newswise.com/articles/ai-safety-institute-launched-as-korea-s-ai-research-hub. Accessed: 17 May 2026; KOREA TIMES. *Korea's AI Safety Institute Officially Launches*. 27 Nov. 2024. Available at:https://www.koreatimes.co.kr/business/tech-science/20241127/koreas-ai-safety-institute-officially-launches. Accessed: 23 May 2026.

61 NATIONAL RESEARCH COUNCIL OF SCIENCE AND TECHNOLOGY. A*I Safety Institute Launched as Korea's AI Research Hub*. Newswise. 27 Nov. 2024. Available at:https://www.newswise.com/articles/ai-safety-institute-launched-as-korea-s-ai-research-hub. Accessed: 17 May 2026.

length from it.

3. FROM AISI TO AISO: THE BRAZILIAN PROPOSAL

In February 2025, the UK renamed its body the *AI Security Institute*, dropping its focus on bias and freedom of speech and refocusing on crime and national security. [62] The acronym was preserved, the noun was not. The shift was substantive: a body that had been founded to address the broad spectrum of risks arising from advanced AI was reframed.

Beyond the UK's specific case, the International AI Safety Report, produced under the chairmanship of Yoshua Bengio and presented at the India summit in 2026, reflects a broader consolidation of this orientation. AI capabilities are acknowledged throughout the report, signalling the importance of understanding and identifying them within the institutional scope of AISIs. Yet they are treated primarily in descriptive terms. Risks, by contrast, occupy the report's analytical and normative core: they are identified, evaluated, and discussed in light of possible mitigations, governance mechanisms, and policy responses.[63]

This asymmetry in the treatment of capabilities and risks is coherent for first-wave jurisdictions: all are developed nations that sit at the frontier of AI development and whose principal stake is mitigating the downside of a technology whose upside they already command.

For a Global South country, the asymmetry sits uneasily, because safety is perceived as necessary but not sufficient. The Indian foreword to the International AI Safety Report 2026 identified, as priorities for India and the Global South, the central

---

[62] McKEON, Christopher. Rebranded AI Security Institute to Drop Focus on Bias and Free Speech. *The Independent*. 14 Feb. 2025. Available at: https://www.independent.co.uk/news/uk/politics/peter-kyle-keir-starmer-rishi-sunak-government-san-francisco-b2698033.html. Accessed: 29 May 2026; WILSON, Claudia. *AI Safety Is Becoming AI Security*. Washington, D.C.: Center for AI Policy. 14 Feb. 2025. Available at: https://www.centeraipolicy.org/work/ai-safety-is-becoming-ai-security. Accessed: 26 May 2026.

[63] INTERNATIONAL AI SAFETY REPORT. *International AI Safety Report 2026*, pp. 10-11, 2026. Available at: https://internationalaisafetyreport.org/publication/international-ai-safety-report-2026. Accessed: 24 May 2026;

issues that the AISO framing seeks to incorporate: bridging the digital divide, ensuring equitable access to capability, and treating AI not only as a source of risk but as a developmental asset whose distribution is itself a policy question.[64] It is within this Global South reality that one can advocate for the construction of AISOs as the norm for the second wave, and within this framework that we can look to Brazil's specific reality in order to understand how an AISO - not an AISI- can (and should) be established in the country.

An AI Safety and Opportunities Institute (AISO) registers the dual stake explicitly. Within it, safety remains within the mandate, since frontier-risk concerns are undoubtedly relevant to a Brazilian audience, but opportunities are added, not as a rhetorical flourish, but as a substantive functional commitment.

4. AN AISO IN BRAZIL: CHARACTERISTICS AND FUNCTIONS IN PRACTICE

To understand the existing infrastructure upon which the Brazilian AISO would be created, it is important to first take note of the national AI plan, the Plano Brasileiro de Inteligência Artificial (PBIA).

Presented in 2024 by the Ministry of Science, Technology and Innovation under the coordination of the National Council for Science and Technology and titled "AI for the Good of All", the national plan foresees investments of R$ 23.03 billion through five structural axes: infrastructure and AI development; diffusion, training and qualification; AI for improving public services; AI for business innovation; and support for regulatory and governance processes.[65]

Its stated objectives include positioning Brazil at the technological frontier, devel-

---

[64] INTERNATIONAL AI SAFETY REPORT. *International AI Safety Report 2026*. 2026. Available at: https://internationalaisafetyreport.org/publication/international-ai-safety-report-2026. Accessed: 24 May 2026; KRISHNAN, Murali. *India's AI Governance Push Takes Center Stage at Summit*, Deutsche Welle, 20 February 2026.

[65] MINISTRY OF SCIENCE, TECHNOLOGY AND INNOVATION (MCTI); CENTRE FOR STRATEGIC STUDIES AND MANAGEMENT (CGEE). *AI for the Good of All: Brazilian Artificial Intelligence Plan*, pp. 9, 26. 2024. Available at: https://pbia.cgee.org.br/. Accessed: 19 May 2026.

oping Large Language Models in Portuguese based on national data, building sovereign cloud infrastructure for the storage of strategic national information, and leveraging Brazil's clean energy matrix to position the country as a global leader in sustainable AI.[66] The plan also identifies immediate impact actions in priority sectors: agricultural extension through AI-supported climate and territorial guidance for rural producers; reduction of school dropout through intelligent attendance management; food quality control for the National School Feeding Programme, which serves approximately forty million students; and a range of further applications in health, public services, and social protection.[67]

The PBIA does not establish a body tasked with evaluating AI systems, developing technical deployment standards, coordinating research across the national AI ecosystem, or representing Brazil within the international network of safety institutes. An AISO would fill this gap, designed to complement the institutional reality the PBIA and other consolidated bodies have created.

For example, the National Institute of Metrology, Quality and Technology (INMETRO) already conducts work on standardisation and, through its Directorate of Scientific Metrology, research into AI safety and reliability, having recently established its Centro de Inteligência Artificial do INMETRO (CIAI). The ANPD (National Data Protection Agency) already addresses AI-related concerns from the perspective of personal data protection[68], and so on.

The institutional design proposed here is one of complementarity, not displacement. An AISO would be conceived to do what the existing organisations cannot do alone, and to support those organisations in doing their own work better. This is the design of an organisation whose incentive is to perform what others cannot perform

---

66 Ibidem.

67 Ibidem, pp. 35-46.

68 INSTITUTO DE TECNOLOGIA E SOCIEDADE DO RIO DE JANEIRO (ITS Rio); CENTRO BRASILEIRO DE RELAÇÕES INTERNACIONAIS (CEBRI). *Um AI Safety and Opportunities Institute para o Brasil: um desenho de características, funções e potenciais desafios*. Rio de Janeiro: ITS Rio; CEBRI, Mar. 2025, p. 6. Available at: https://itsrio.org/en/publicacoes/an-ai-safety-and-opportunities-institute-for-brazil-a-design-of-features-functions-and-potential-challenges/. Accessed: 17 May 2026.

on their own, while assisting the others in fulfilling their core functions.

Here the comparative analysis developed in earlier sections becomes useful in a different way. The first-wave AISIs were built around a specific developmental position. The UK, the USA, Japan, Singapore, Canada, France, South Korea, and the EU are all jurisdictions that occupy, in different ways, the upstream end of the global AI value chain. They host the leading frontier developers, or they are positioned to evaluate the products of those developers before deployment, or they exercise regulatory authority over the firms that produce them. Their AISIs reflect this position. The principal concern is the mitigation of risks arising from technologies whose capabilities the same jurisdictions are also shaping. A body that has pre-deployment access to OpenAI and Anthropic, as the UK AISI does, can be technical and advisory rather than regulatory, because the underlying access already exists.[69]

Brazil is not in that position. The country does not yet host firms developing frontier general-purpose AI models, and the empirical evidence indicates that the gap between countries that do and do not is widening rather than closing. The International AI Safety Report 2025 documents that the cost of training state-of-the-art AI models has grown by a factor of two to three per year over the past eight years and could exceed USD 1 billion per model by 2027.[70] The same report identifies an AI R&D divide between high-income economies and lower-middle-income countries, driven by uneven distribution of devices, talent, financial resources, and digital

---

[69] ARAUJO, Renan; FORT, Kristina; GUEST, Oliver. *Understanding the First Wave of AI Safety Institutes: Characteristics, Functions, and Challenges*. Oxford: Institute for AI Policy and Strategy (IAPS), p. 19, 35-38. 2024. Available at: https://doi.org/10.48550/arXiv.2410.09219. Accessed: 22 May 2026. 2024; INSTITUTO DE TECNOLOGIA E SOCIEDADE DO RIO DE JANEIRO (ITS Rio); CENTRO BRASILEIRO DE RELAÇÕES INTERNACIONAIS (CEBRI). *Um AI Safety and Opportunities Institute para o Brasil: um desenho de características, funções e potenciais desafios*. Rio de Janeiro: ITS Rio; CEBRI, Mar. 2025, pp. 9, 17, 35. Available at: https://itsrio.org/en/publicacoes/an-ai-safety-and-opportunities-institute-for-brazil-a-design-of-features-functions-and-potential-challenges/. Accessed: 27 May 2026; AI SECURITY INSTITUTE (AISI). *About the AI Security Institute*. London: UK Government. 2024. Available at: https://www.aisi.gov.uk/about. Accessed: 14 May 2026.

[70] INTERNATIONAL AI SAFETY REPORT. *International AI Safety Report 2025*, p. 93. 2025. Available at: https://internationalaisafetyreport.org/publication/international-ai-safety-report-2025. Accessed: 22 May 2026.

infrastructure, and notes that academic institutions and most companies in such countries lack the means to compete with the largest technology firms.[71] India has succeeded in increasing its concentration of skilled AI talent by 263 per cent since 2016, more than any other major economy in that period, demonstrating that the divide can be partially addressed when the institutional incentives align.[72] But the broader trajectory is one of widening dispersion.

The Brazilian position within this divide is empirically documented. The *Mapa da GenAI no Brasil* (Brazil's GenAI Map), surveyed 350 executives across Brazilian organisations of varied size and sector. It found that 73% of organisations already discuss GenAI in decision-making forums, while only 16% have a formal strategy implemented.[73] 46% of organisations using GenAI have no defined ethical guidelines for its use. 8% of organisations have zero training capacity in the technology, and a further 35% report low capacity.[74]

Overall, most local GenAI applications in Brazil depend on models and infrastructures developed elsewhere, leaving the country, as Andre Miceli of MIT Technology Review Brasil observed, "more vulnerable from a strategic and economic standpoint".[75] Therefore, the country is currently predominantly a consumer of AI capabilities developed abroad, with strategic discussion that has not yet translated into structured institutional response. The diagnosis matters because it speaks directly to the question of national sovereignty: positioning Brazil as a producer of technology rather than merely a consumer of solutions developed elsewhere is one of the explicit motivations for the creation of an AISO.[76]

---

71 Ibidem, p. 119.

72 Ibidem, p. 45.

73 TEC.INSTITUTE; MIT TECHNOLOGY REVIEW BRASIL. *O Mapa da GenAI no Brasil: Da Estratégia à Implementação*, pp. 9-10. 2025. Available at: https://rd.mittechreview.com.br/tec-report_genai. Accessed: 28 May 2026.

74 Ibidem, pp. 11-12.

75 Ibidem, p. 16.

76 INSTITUTO DE TECNOLOGIA E SOCIEDADE DO RIO DE JANEIRO (ITS Rio); CENTRO BRASILEIRO DE RELAÇÕES INTERNACIONAIS (CEBRI). *Um AI Safety and Opportunities Institute para o Brasil: um desenho de características, funções e potenciais desafios*. Rio de Janeiro: ITS Rio; CEBRI, Mar. 2025, p. 43. Available at: https://itsrio.org/en/publicacoes/an-ai-safety-and-opportunities-institute-for-brazil-a-design-of-features-functions-and-potential-challenges/.

The picture is complicated further by the underlying conditions of digital access. The Brazilian National Telecommunications Agency, ANATEL, published in September 2025 a national survey on meaningful connectivity, conducted in partnership with the Brazilian Institute for Consumer Defence (IDEC) and the International Telecommunication Union (ITU).[77] The framework adopted defines meaningful connectivity not as the binary presence of an internet connection but as "the level of connectivity that allows everyone to access the internet with a safe, satisfactory and enriching online experience, including affordable cost of the technology."[78]

The ANATEL survey found that 12% of Brazilians with monthly income of up to one minimum wage went more than fifteen days without mobile internet access due to data plan exhaustion, in the thirty days preceding the interview. When it comes to devices, it found that 30% of the population reported not owning a computer, whether because they did not know how to use one or had no interest in doing so.[79]

In a country whose population's relationship with the digital infrastructure is uneven in measurable ways, the policy question raised by AI is not only how to prevent catastrophic misuse of frontier models. It is also how to ensure that the technology reaches and effectively produces value for the population. Therefore, it is important to understand the format that this structure would need to have in order to fulfill these objectives.

4.1 Characteristics

The first wave of AISIs converged around five relatively stable institutional characteristics: (i) institutional autonomy combined with technical and scientific open-

Accessed: 27 May 2026.

77 ANATEL. *Pesquisa de Conectividade Significativa*, pp. 3-4. Superintendência de Relações com Consumidores. Sept. 2025. Available at: https://sei.anatel.gov.br/sei/modulos/pesquisa/md_pesq_documento_consulta_externa.php?8-74Kn1tDR89f1Q7RjX8EYU46IzCFD26Q9Xx5QNDbqZUQ5y5j6h2yvy7aYmCm-1MuqUgXUpnTgbbGy-Xr2hDZz2dxAx3AxqvxAyuwr6onoLCKa82vajrfdk0dWk5CwV. Accessed: 25 May 2026.

78 Ibidem, p. 3.

79 Ibidem, pp. 19-33.

ness, (ii) expert-led and evidence-based operation, (iii) a mandate restricted to advanced AI systems, (iv) the absence of direct regulatory powers, and (v) participation in an international network of reciprocal cooperation.[80] The Brazilian AISO proposed here preserves these core characteristics, but reformulates them in light of the "opportunities" commitment introduced earlier and of the conditions identified in the analysis of Brazilian institutional infrastructure.[81]

The first characteristic concerns institutional independence. The Paris AI Action Summit demonstrated, more clearly than previous summits, the extent to which institutions subordinated to a single political authority remain vulnerable to rapid political reorientation. The AISI's transition from "safety" to "security", as well as the reconfiguration of the USA AISI following the change in administration, illustrate how institutional continuity may depend excessively on Federal Government priorities. A Brazilian AISO designed to withstand political discontinuity should therefore not operate merely as a department within a single ministry.

Brazil has a track record of multistakeholder governance tradition that has been developed within internet governance, with ICANN at the international level and the *Comitê Gestor da Internet no Brasil* at the domestic level. These arrangements have historically functioned as coordination environments capable of reducing informational asymmetries, distributing institutional authority, and limiting capture by any single public or private actor.[82] The proposed AISO would follow a similar logic:

---

[80] ARAUJO, Renan; FORT, Kristina; GUEST, Oliver. *Understanding the First Wave of AI Safety Institutes: Characteristics, Functions, and Challenges*. Oxford: Institute for AI Policy and Strategy (IAPS), p.1. 2024. Available at: https://doi.org/10.48550/arXiv.2410.09219. Accessed: 22 May 2026; INSTITUTO DE TECNOLOGIA E SOCIEDADE DO RIO DE JANEIRO (ITS Rio); CENTRO BRASILEIRO DE RELAÇÕES INTERNACIONAIS (CEBRI). *Um AI Safety and Opportunities Institute para o Brasil: um desenho de características, funções e potenciais desafios*. Rio de Janeiro: ITS Rio; CEBRI, Mar. 2025, p. 7. Available at: https://itsrio.org/en/publicacoes/an-ai-safety-and-opportunities-institute-for-brazil-a-design-of-features-functions-and-potential-challenges/. Accessed: 20 May 2026.

[81] It is important to note that the AISO's success will depend not only on its institutional design but also on the political and financial conditions of its implementation.

[82] INSTITUTO DE TECNOLOGIA E SOCIEDADE DO RIO DE JANEIRO (ITS Rio); CENTRO BRASILEIRO DE RELAÇÕES INTERNACIONAIS (CEBRI). *Um AI Safety and Opportunities Institute para o Brasil: um desenho de características, funções e potenciais desafios*. Rio de Janeiro:

institutionally anchored within the federal government, but administered through a multistakeholder governance structure composed of representatives from government, academia, civil society and the technical sector. Such a structure would not eliminate political pressures, but it would increase institutional continuity and preserve the informational openness necessary for technical cooperation and risk assessment.

The second characteristic concerns expert-led governance grounded in technical evidence. In the Brazilian context, however, the relevant expertise cannot be restricted to frontier-model evaluation and systemic-risk assessment alone. Unlike the jurisdictions in which the first AISIs emerged, Brazil does not occupy a central position in the global AI production chain. The institutional challenges associated with advanced AI therefore include not only the mitigation of potential harms, but also the identification of developmental bottlenecks capable of deepening technological dependency.

For this reason, the AISO would require multidisciplinary expertise extending beyond safety in the narrow sense. Alongside capabilities in model evaluation, cybersecurity, and technical auditing, the institution would also need expertise related to computational infrastructure, semiconductor dependency, public digital infrastructure, industrial policy, scientific capacity-building, and technological diffusion. This broader evidentiary structure would allow the institution to produce national capability assessments, identify strategic infrastructural vulnerabilities, and evaluate how global AI transformations affect domestic technological autonomy. The integration of these different forms of expertise is institutionally coherent with the multistakeholder structure proposed above, since the challenges faced by peripheral economies involve coordination problems that exceed the boundaries of technical safety alone.

The third characteristic concerns the consultative and non-regulatory nature of the institution. The Brazilian AISO would preserve this characteristic strictly. Its purpose would not be to replace or overlap existing regulatory authorities, but to function as a technical coordination and advisory body capable of supporting them.

ITS Rio; CEBRI, Mar. 2025, p. 33. Available at: https://itsrio.org/en/publicacoes/an-ai-safety-and-opportunities-institute-for-brazil-a-design-of-features-functions-and-potential-challenges/. Accessed: 22 May 2026.

This design is particularly important in the Brazilian context, where the institutional debates surrounding Bill No. 2338/2023 increasingly indicate the possible consolidation of a distributed governance architecture involving multiple sectoral authorities. Within such a framework, the creation of an additional regulator could contribute to institutional fragmentation and jurisdictional conflict rather than alleviate them.

The absence of direct enforcement powers would therefore constitute an institutional advantage rather than a limitation. By remaining outside the sphere of coercive regulation, the AISO could operate as a trusted environment for technical cooperation, information-sharing, and methodological development across institutions that currently possess fragmented forms of expertise. Its role would be to support public authorities through technical studies, shared evaluation protocols, risk methodologies, incident analysis, and cross-sector coordination mechanisms, while preserving the autonomy of existing regulators.

The fourth characteristic concerns the AISO's function as a coordinating node within a broader institutional network. The institution should not be conceived as a centralized regulator, but rather as a " centre of centres" capable of articulating communication among already existing public authorities, research institutions, technical bodies, and international partners.

The international dimension of this networked structure is equally important. The Brazilian AISO should participate actively in the emerging international ecosystem of AI governance, not merely as a recipient of foreign standards, but as a contributor to their development. Existing international discussions remain heavily concentrated in jurisdictions positioned at the centre of AI production and computational infrastructure. Brazil, however, confronts governance challenges associated with technological dependency, infrastructural asymmetry, and uneven digital development that remain underrepresented in current global frameworks.

Participation in the international AISO network would therefore allow Brazil not only to absorb external methodologies, but also to contribute perspectives derived from Global South institutional realities, including experiences with multistakeholder digital governance, public digital infrastructure, and governance under conditions of infrastructural scarcity. A reciprocal model of cooperation of this kind would strengthen Brazil's ability to participate in international AI governance debates as an

active institutional interlocutor rather than as a peripheral rule-taker.

4.2 Functions

AISIs typically deal with a functional triad of research, development of standards, and cooperation.[83] Regarding research, the intended posture is empirical and problem-oriented, aimed at advancing "the science of AI safety and opportunities", and at expanding the technical knowledge needed to make AI systems safer and to broaden the conditions under which AI can be productively deployed.[84] The activities recognisable from the first wave fall within this function in the safety register: safety evaluations of frontier models, assessment of potential and emerging risks of critical AI systems, provision of evidence to policy-making processes, advancement of foundational AI safety research, and the development of testing mechanisms and methodologies for the safe deployment of AI models, systems and applications.[85]

---

[83] ARAUJO, Renan; FORT, Kristina; GUEST, Oliver. *Understanding the First Wave of AI Safety Institutes: Characteristics, Functions, and Challenges*. Oxford: Institute for AI Policy and Strategy (IAPS), pp. 17-30. 2024. Available at: https://doi.org/10.48550/arXiv.2410.09219. Accessed: 22 May 2026; INSTITUTO DE TECNOLOGIA E SOCIEDADE DO RIO DE JANEIRO (ITS Rio); CENTRO BRASILEIRO DE RELAÇÕES INTERNACIONAIS (CEBRI). *Um AI Safety and Opportunities Institute para o Brasil: um desenho de características, funções e potenciais desafios*. Rio de Janeiro: ITS Rio; CEBRI, Mar. 2025, p. 39. Available at: https://itsrio.org/en/publicacoes/an-ai-safety-and-opportunities-institute-for-brazil-a-design-of-features-functions-and-potential-challenges/. Accessed: 21 May 2026.

[84] INSTITUTO DE TECNOLOGIA E SOCIEDADE DO RIO DE JANEIRO (ITS Rio); CENTRO BRASILEIRO DE RELAÇÕES INTERNACIONAIS (CEBRI). *Um AI Safety and Opportunities Institute para o Brasil: um desenho de características, funções e potenciais desafios*. Rio de Janeiro: ITS Rio; CEBRI, Mar. 2025, p. 39. Available at: https://itsrio.org/en/publicacoes/an-ai-safety-and-opportunities-institute-for-brazil-a-design-of-features-functions-and-potential-challenges/. Accessed: 27 May 2026.

[85] INSTITUTO DE TECNOLOGIA E SOCIEDADE DO RIO DE JANEIRO (ITS Rio); CENTRO BRASILEIRO DE RELAÇÕES INTERNACIONAIS (CEBRI). *Um AI Safety and Opportunities Institute para o Brasil: um desenho de características, funções e potenciais desafios*. Rio de Janeiro: ITS Rio; CEBRI, Mar. 2025, p. 39. Available at: https://itsrio.org/en/publicacoes/an-ai-safety-and-opportunities-institute-for-brazil-a-design-of-features-functions-and-potential-challenges/. Accessed: 27 May 2026; ARAUJO, Renan; FORT, Kristina; GUEST, Oliver. *Understanding the First Wave of AI Safety Institutes: Characteristics, Functions, and Challenges*. Oxford: Institute for

The Brazilian translation of this function would preserve all of these activities, but would add to them a parallel research stream concerned with opportunities: empirical evaluation of how AI is being deployed in Brazilian healthcare, education, agriculture, public services and labour markets; assessment of which deployments produce measurable improvements and which do not; identification of methodologies that allow AI capabilities developed abroad to be productively adapted to Brazilian linguistic, cultural and institutional contexts, among others. The research function would thus operate in two registers simultaneously, with the same analytical rigour applied to both. Its outputs would include not only evaluations of frontier risk but also evidence-based guidance for the deployment of AI in domains where the Brazilian state has substantial responsibility and where the population's exposure to AI systems is already significant.

The research function also carries implications for how research is produced. The first-wave AISIs have made extensive use of collaborative research arrangements with industry, academia, civil society, and peer governments; have established formal partnerships; have opened calls for research proposals; and have provided funding for fellowships.[86] The Brazilian AISO would inherit these mechanisms and adapt them to local conditions, prioritising research on questions that are structurally specific to Brazil's linguistic, economic, and institutional context.

The second function is the development of standards and guided practices. The aim is to influence AI development and deployment through applied contributions such as technical guidelines, evaluation protocols, and standardised methodologies.[87] First-wave examples include the development of technical safety guidelines and

AI Policy and Strategy (IAPS), pp. 17-22. 2024. Available at: https://doi.org/10.48550/arXiv.2410.09219. Accessed: 22 May 2026.

[86] INSTITUTO DE TECNOLOGIA E SOCIEDADE DO RIO DE JANEIRO (ITS Rio); CENTRO BRASILEIRO DE RELAÇÕES INTERNACIONAIS (CEBRI). *Um AI Safety and Opportunities Institute para o Brasil: um desenho de características, funções e potenciais desafios*. Rio de Janeiro: ITS Rio; CEBRI, Mar. 2025, p. 39. Available at: https://itsrio.org/en/publicacoes/an-ai-safety-and-opportunities-institute-for-brazil-a-design-of-features-functions-and-potential-challenges/ . Accessed: 27 May 2026.

[87] INSTITUTO DE TECNOLOGIA E SOCIEDADE DO RIO DE JANEIRO (ITS Rio); CENTRO BRASILEIRO DE RELAÇÕES INTERNACIONAIS (CEBRI). *Um AI Safety and Opportunities*

tools, and software libraries to operationalise these methodologies.[88] The Brazilian AISO would carry out this function in close articulation with INMETRO and its already established conformity assessment, working to contribute to the methodological foundation on which AI-specific conformity assessment could be developed, harmonised with international standards, and adapted to the conditions of the Brazilian market.[89]

Lastly, regarding cooperation, the aim is to bridge interaction gaps within the ecosystem and to promote international alignment, with cooperation conceived across three dimensions: multisectoral cooperation, facilitated interaction between market actors, and international exchange.[90] The Brazilian translation of this function carries a specific dimension worth emphasising as the country would not only follow global AI safety standards but also share its own solutions, frameworks, and perspectives with other jurisdictions, strengthening Brazil's position as an active participant in international discussions and amplifying its capacity to influence the global agenda while simultaneously incorporating best practices developed elsewhere[91], creating an institutionally significant reciprocal posture.

Taken together, these three functions describe an institution that operates simultaneously as a producer of evidence, a developer of standards, and a convener of

---

Institute para o Brasil: um desenho de características, funções e potenciais desafios. Rio de Janeiro: ITS Rio; CEBRI, Mar. 2025, pp. 39-40. Available at: https://itsrio.org/en/publicacoes/an-ai-safety-and-opportunities-institute-for-brazil-a-design-of-features-functions-and-potential-challenges/. Accessed: 27 May 2026; ARAUJO, Renan; FORT, Kristina; GUEST, Oliver. *Understanding the First Wave of AI Safety Institutes: Characteristics, Functions, and Challenges*. Oxford: Institute for AI Policy and Strategy (IAPS), pp. 25-29. 2024. Available at: https://doi.org/10.48550/arXiv.2410.09219. Accessed: 22 May 2026.

88 INSTITUTO DE TECNOLOGIA E SOCIEDADE DO RIO DE JANEIRO (ITS Rio); CENTRO BRASILEIRO DE RELAÇÕES INTERNACIONAIS (CEBRI). *Um AI Safety and Opportunities Institute para o Brasil: um desenho de características, funções e potenciais desafios*. Rio de Janeiro: ITS Rio; CEBRI, Mar. 2025, pp. 39-40. Available at: https://itsrio.org/en/publicacoes/an-ai-safety-and-opportunities-institute-for-brazil-a-design-of-features-functions-and-potential-challenges/ Accessed: 27 May 2026.

89 Ibidem, pp. 16-42.

90 Ibidem, p. 40.

91 Ibidem, p. 37.

cooperation. They are mutually reinforcing: credible research strengthens standard-setting, and cooperative relationships expand the scope of both. This is what the Brazilian AISO should strive to achieve.

5. CONCLUSION

This paper has examined the first wave of AISIs, mapped the structural conditions under which they were built, and argued why those conditions cannot be unalterably transplanted to Brazil. The first-wave model worked because the countries that built it were, in different ways, already upstream: they hosted the developers, commanded the infrastructure, and held the access. Their principal governance challenge was to manage the risks of a technology whose benefits they were already capturing.

Brazil does not share that position. A body modelled on first-generation AISIs are technically rigorous, internationally networked, focused on the empirical evaluation of frontier systems. It would be a genuine contribution to Brazilian AI governance, but ignore AI governance questions concern deployment, access, linguistic and cultural adaptation, and the conditions under which a technology developed elsewhere can be made to serve a population of two hundred million people with uneven digital infrastructure and heterogeneous institutional capacity.

The AISO model is a response to that incompleteness. It preserves what is valuable in the first-wave template: the technical grounding, the advisory posture, and the connection to an international reciprocal network; while also extending its mandate to encompass opportunities as a substantive analytical commitment. Safety and opportunity are treated not as competing priorities requiring political negotiation, but as two registers of the same empirical question: what does this technology do, to whom, under what conditions, and with what consequences? An institution capable of asking that question rigorously, and of producing evidence that can guide both regulation and deployment, is one that Brazil does not currently have and that the existing institutional landscape (e.g. PBIA, INMETRO, ANPD, CGI.br) cannot supply alone.

The broader significance of this proposal extends beyond Brazil. The trajectory of the international summit process from Bletchley's safety vocabulary through Seoul's inclusivity agenda to Paris's emphasis on action and New Delhi's reframing around

impact suggests that the second wave of AI governance institutions will not be built in the image of the first. The countries that will shape that wave are precisely those, like Brazil, whose stake in AI governance is simultaneously defensive and developmental. An AISO is not a departure from the international consensus; it is a contribution to the next iteration of it. Brazil arrives at this moment with the diplomatic standing, the institutional foundations, and the analytical clarity to lead that iteration rather than merely follow it.

BIBLIOGRAPHY

AI SECURITY INSTITUTE (AISI). *About the AI Security Institute*. London: UK Government. 2024. Available at: https://www.aisi.gov.uk/about. Accessed: 14 May 2026.

ALLEN, Gregory C.; ADAMSON, Georgia. *The AI Safety Institute International Network: Next Steps and Recommendations*. Washington, D.C.: Center for Strategic and International Studies (CSIS). 2024. Available at: https://www.csis.org/analysis/ai-safety-institute-international-network-next-steps-and-recommendations. Accessed: 19 May 2026.

ANATEL. *Pesquisa de Conectividade Significativa*. Superintendência de Relações com Consumidores. Sept. 2025. Available at: https://sei.anatel.gov.br/sei/modulos/pesquisa/md_pesq_documento_consulta_externa.php?8-74Kn1tDR89f1Q7RjX8EYU46IzCFD26Q9Xx5QNDbqZUQ5y5j6h2yvy7aYmCm-1MuqUgXUpnTgbbGy-Xr2hDZz2dxAx3AxqvxAyuwr6onoLCKa82vajrfdk0dWk5CwV. Accessed: 25 May 2026.

ARAUJO, Renan; FORT, Kristina; GUEST, Oliver. *Understanding the First Wave of AI Safety Institutes: Characteristics, Functions, and Challenges*. Oxford: Institute for AI Policy and Strategy (IAPS). Oct. 2024. Available at: https://doi.org/10.48550/arXiv.2410.09219. Accessed: 22 May 2026.

CANADIAN INSTITUTE FOR ADVANCED RESEARCH (CIFAR). *CIFAR to Lead an AI Safety Research Program in Support of the Canadian AI Safety Institute*. 12 Nov. 2024. Available at: https://cifar.ca/cifarnews/2024/11/12/government-of-canada-announces-canadian-ai-safety-institute/. Accessed: 22 May 2026.

CHAUDHURI, Rudra. *Disrupting AI Safety Institutes: The India Way*. Washington, D.C.: Carnegie Endowment for International Peace. Sept. 2024. Available at: https://carnegieendowment.org/posts/2024/09/disrupting-ai-safety-institutes-the-india-way. Accessed: 24 May 2026.

CUÉLLAR, Mariano-Florentino. *The UK AI Safety Summit Opened a New Chapter in AI Diplomacy*. Washington, D.C.: Carnegie Endowment for International Peace. Nov. 2023. Available at: https://carnegieendowment.org/posts/2023/11/the-uk-ai-safety-summit-opened-a-new-chap-

ter-in-ai-diplomacy. Accessed: 20 May 2026.

DIGITAL TRUST CENTRE (DTC). *About Us*. Singapore: Nanyang Technological University (NTU). 2024. Available at: https://www.ntu.edu.sg/dtc/about-us. Accessed: 16 May 2026.

ÉLYSÉE PALACE. *AI Action Summit*. Paris. Feb. 2025. Available at: https://www.elysee.fr/en/sommet-pour-l-action-sur-l-ia. Accessed: 18 May 2026.

EUROPEAN COMMISSION. *European AI Office: Shaping Europe's Digital Future*. Brussels. 2024. Available at: https://digital-strategy.ec.europa.eu/en/policies/ai-office. Accessed: 13 May 2026.

HERN, Alex. *Trying to Tame AI: Seoul Summit Flags Hurdles to Regulation*. The Guardian. 27 May 2024. Available at: https://www.theguardian.com/technology/article/2024/may/27/trying-to-tame-ai-seoul-summit-flags-hurdles-to-regulation. Accessed: 23 May 2026.

INNOVATION, SCIENCE AND ECONOMIC DEVELOPMENT CANADA. *Canada Launches Canadian Artificial Intelligence Safety Institute*. 12 Nov. 2024. Available at: https://www.canada.ca/en/innovation-science-economic-development/news/2024/11/canada-launches-canadian-artificial-intelligence-safety-institute.html. Accessed: 20 May 2026.

INSTITUTO DE TECNOLOGIA E SOCIEDADE DO RIO DE JANEIRO (ITS Rio); CENTRO BRASILEIRO DE RELAÇÕES INTERNACIONAIS (CEBRI). *Um AI Safety and Opportunities Institute para o Brasil: um desenho de características, funções e potenciais desafios*. Rio de Janeiro: ITS Rio; CEBRI. Mar. 2025. Available at: https://itsrio.org/en/publicacoes/an-ai-safety-and-opportunities-institute-for-brazil-a-design-of-features-functions-and-potential-challenges/. Accessed: 14 May 2026.

INTERNATIONAL AI SAFETY REPORT. *International AI Safety Report 2025*. 2025. Available at: https://internationalaisafetyreport.org/publication/international-ai-safety-report-2025. Accessed: 22 May 2026.

INTERNATIONAL AI SAFETY REPORT. *International AI Safety Report 2026*. 2026. Available at: https://internationalaisafetyreport.org/publication/international-ai-safety-report-2026. Accessed: 24 May 2026.

INTERNATIONAL NETWORK OF AI SAFETY INSTITUTES. *Mission Statement*. San Francisco. 21 Nov. 2024. Available at: https://www.nist.gov/system/files/documents/2024/11/20/Mission%20Statement%20-%20International%20Network%20of%20AISIs.pdf. Accessed: 17 May 2026.

KOREA TIMES. *Korea's AI Safety Institute Officially Launches*. 27 Nov. 2024. Available at: https://www.koreatimes.co.kr/business/tech-science/20241127/koreas-ai-safety-institute-officially-launches. Accessed: 23 May 2026.

KRISHNAN, Murali. *India's AI Governance Push Takes Center Stage at Summit*. Deutsche Welle. 20 Feb. 2026. Available at: https://www.dw.com/en/indias-ai-governance-push-takes-center-stage-at-summit/a-76058826. Accessed: 27 May 2026.

LEONE, Daniela et al. *Evaluating the Potential Functions of an International Institution for AI Safety: An Institutional Analysis of the IAEA and IPCC, in the Context of Recent Trends in AI Safety*. London: The Alan Turing Institute. arXiv:2409.10536v1. Sept. 2024. Available at: https://doi.org/10.48550/arXiv.2409.10536. Accessed: 26 May 2026.

McKEON, Christopher. *Rebranded AI Security Institute to Drop Focus on Bias and Free Speech*. The Independent. 14 Feb. 2025. Available at: https://www.independent.co.uk/news/uk/politics/peter-kyle-keir-starmer-rishi-sunak-government-san-francisco-b2698033.html. Accessed: 29 May 2026.

MINISTÈRE DE L'ÉCONOMIE ET DES FINANCES. *La France se dote d'un Institut national pour l'évaluation et la sécurité de l'intelligence artificielle (INESIA)*. Jan. 2025. Available at: https://www.economie.gouv.fr/actualites/la-france-se-dote-dun-institut-national-pour-levaluation-et-la-securite-de-lintelligence#. Accessed: 15 May 2026.

MINISTRY OF ECONOMY, TRADE AND INDUSTRY (METI). *Launch of AI Safety Institute*. 14 Feb. 2024. Available at: https://www.meti.go.jp/english/press/2024/0214_001.html. Accessed: 18 May 2026.

MINISTRY OF SCIENCE, TECHNOLOGY AND INNOVATION (MCTI); CENTRE FOR STRATEGIC STUDIES AND MANAGEMENT (CGEE). *AI for the Good of All: Brazilian Artificial Intelligence Plan*. 2024. Available at: https://pbia.cgee.org.br/. Accessed: 19 May 2026.

NATIONAL RESEARCH COUNCIL OF SCIENCE AND TECHNOLOGY. *AI Safety Institute Launched as Korea's AI Research Hub*. Newswise. 27 Nov. 2024. Available at: https://www.newswise.com/articles/ai-safety-institute-launched-as-korea-s-ai-research-hub. Accessed: 17 May 2026.

PERRIGO, Billy. *U.S. Pushes for Less AI Regulation at Paris Summit*. TIME. Feb. 2025. Available at: https://time.com/7221384/ai-regulation-takes-bac. Accessed: 21 May 2026.

POUGET, Hadrien. *France's AI Summit Is a Chance to Reshape Global Narratives on AI*. Washington, D.C.: Carnegie Endowment for International Peace. July 2024. Available at: https://carnegieendowment.org/posts/2024/07/france-ai-summit-reshape-global-narrative. Accessed: 21 May 2026.

TEC.INSTITUTE; MIT TECHNOLOGY REVIEW BRASIL. *O Mapa da GenAI no Brasil: Da Estratégia à Implementação*. 2025. Available at: https://rd.mittechreview.com.br/tec-report_genai. Accessed: 28 May 2026.

THURNHERR, Lara. *Which Information should the UK and US AISI share with an International Network of AISIs?* London: King's College London. Mar. 2025. arXiv:2503.04741. Available at: https://arxiv.org/abs/2503.04741. Accessed: 27 May 2026.

UK GOVERNMENT. *About the AI Safety Summit 2023*. GOV.UK. 2023. Available at: https://www.gov.uk/government/topical-events/ai-safety-summit-2023/about. Accessed: 13 May 2026.

UK GOVERNMENT. *About the AI Seoul Summit 2024*. GOV.UK. 2024. Available at: https://www.gov.uk/government/topical-events/ai-seoul-summit-2024/about. Accessed: 11 May 2026.

UK GOVERNMENT. *Countries Agree to Safe and Responsible Development of Frontier AI in Landmark Bletchley Declaration*. GOV.UK. 1 Nov. 2023. Available at: https://www.gov.uk/government/news/countries-agree-to-safe-and-responsible-development-of-frontier-ai-in-landmark-bletchley-declaration. Accessed: 15 May 2026.

UK GOVERNMENT. *Introducing the AI Safety Institute*. Nov. 2023. Available at: https://assets.publishing.service.gov.uk/media/65438d159e05fd0014be7bd9/introducing-ai-safety-institute-web-accessible.pdf. Accessed: 12 May 2026.

UK GOVERNMENT. *Seoul Declaration for Safe, Innovative and Inclusive AI*. AI Seoul Summit. 21 May 2024. Available at: https://www.gov.uk/government/publications/seoul-declaration-for-safe-innovative-and-inclusive-ai-ai-seoul-summit-2024. Accessed: 28 May 2026.

UK GOVERNMENT. *Seoul Statement of Intent Toward International Cooperation on AI Safety Science*. AI Seoul Summit 2024. GOV.UK. 2024. Available at: https://www.gov.uk/government/publications/seoul-declaration-for-safe-innovative-and-inclusive-ai-ai-seoul-summit-2024/seoul-statement-of-intent-toward-international-cooperation-on-ai-safety-science-ai-seoul-summit-2024-annex. Accessed: 17 May 2026.

UK GOVERNMENT. *The Bletchley Declaration by Countries Attending the AI Safety Summit, 1-2 November 2023*. GOV.UK. 1-2 Nov. 2023. Available at: https://www.gov.uk/government/publications/ai-safety-summit-2023-the-bletchley-declaration/the-bletchley-declaration-by-countries-attending-the-ai-safety-summit-1-2-november-2023. Accessed: 16 May 2026.

WILSON, Claudia. *AI Safety Is Becoming AI Security*. Washington, D.C.: Center for AI Policy. 14 Feb. 2025. Available at: https://www.centeraipolicy.org/work/ai-safety-is-becoming-ai-security. Accessed: 26 May 2026.

AI REGULATORY SANDBOXES BEYOND THE HYPE: COMPARATIVE DESIGN CHOICES, GOVERNANCE, AND METRICS FOR ACCOUNTABILITY

Thiago Guimarães de Barros Cobra

Master of Laws (LL.M.) in Administrative Law, University of São Paulo Law School (USP). Specialist in Civil Procedural Law, Escola Superior da Advocacia (ESA); and in Administrative Law and Constitutional Law, Escola Paulista de Direito. MBA in Public and Government Policy Management, Escola Paulista de Direito. LL.B., Mackenzie Presbyterian University. Attorney in Regulatory Law at HGC Advogados

DOI: https://doi.org/10.59224/dti8.ch6

Abstract: This article analyzes AI regulatory sandboxes as instruments of experimental governance and asks when they are legitimate learning devices versus regulatory shortcuts. It argues legitimacy depends on a finite, evidence-driven cycle: defined scope, risk-based safeguards, continuous supervision, auditable metrics (technical and socioeconomic, incl. bias and profiling), and enforceable exit/transition rules. It maps recurring failures, such as perpetuation, arbitrage/sandbox shopping, opacity/capture, harm socialization, and derives design conditions to close cycles on verifiable criteria.

Keywords: regulatory sandboxes; artificial intelligence; experimental regulation

Table of contents: 1. Introduction; 2. Theoretical framework: regulation, risk, and institutional learning; 2.1. Regulating emerging technologies and the limits of the prescriptive model; 2.2. Risk-based regulation: promise, blind spots, and systemic risk; 2.3. Responsive and experimental regulation: institutional learning as a regulatory strategy; 3. International normative backdrop: the foundations of "trustworthy

AI"; 3.1. European Union: from "trustworthy AI" to normative densification; 3.2. United Kingdom: sectoral coordination, and adaptability instruments; 3.3. United States: institutional fragmentation, and political volatility; 3.4. Asian experiences (China, South Korea, and India): state-centered models; 3.5. Digital sovereignty: data, infrastructure, and governance tensions; 4. Institutional design of AI sandboxes: governance, metrics, and limits; 4.1. Purposes of AI sandboxes: regulatory learning and supervised compliance; 4.2. Eligibility and selection: scope boundaries and risk-proportionate design; 4.3. Safeguards and risk management throughout the experimental cycle; 4.4. Governance, continuous oversight, and incident documentation; 4.5. Exit rules and institutionalizing learning: from testing to decision; 5. Metrics, evidence, and the evaluation of experimentalism in AI; 5.1. Performance, robustness, and operational security: minimum evidence for a reliable AI test; 5.2. Socioeconomic impacts and distributive effects: benefits, costs, and externalities of the experiment; 5.3. Evidence governance and cycle integrity: reporting, auditability, and comparability of results; 6. When the sandbox becomes a dangerous "shortcut": risks and conditions of legitimacy; 6.1. Exceptionalism becomes the rule: cycle perpetuation and erosion of exit rules; 6.2. Competitive asymmetry and "sandbox shopping": regulatory arbitrage and market distortions; 6.3. Opacity and capture: informational asymmetry and the weakening of public learning; Conclusion; References.

1. INTRODUCTION

The fast diffusion of artificial intelligence (AI) systems, especially general-purpose models and applications capable of shaping decisions at scale, brings renewed urgency to a classic problem of regulatory law: how to reconcile innovation and technological dynamism with legal certainty, the protection of rights, and market integrity. The difficulty is not merely whether to regulate, but how to regulate when (i) technological cycles accelerate, (ii) impacts vary significantly across contexts of use, and (iii) a substantial portion of the relevant technical and organizational knowledge about system performance, limitations, and risks is concentrated in private actors. This environment intensifies the *pacing problem*: traditional legal instruments often arrive too late, while immediate responses that are overly open-ended or indeterminate

may fail to establish minimum standards of trustworthiness.

Against this backdrop, contemporary regulatory discourse increasingly emphasizes risk-based approaches, lifecycle governance, and supervised experimentation mechanisms. Among these, regulatory sandboxes have gained prominence: they are framed as controlled environments in which innovations can be tested under regulatory supervision, with the stated aims of reducing uncertainty, generating evidence, and feeding regulatory learning. In AI, however, this promise operates within a more complex institutional setting: AI is cross-cutting, affects multiple sectors, and often involves fragmented value chains (model providers, integrators, distributors, professional users, and end users), which raises coordination costs and complicates the allocation of responsibilities.

The very flexibility that makes sandboxes attractive can produce counterproductive outcomes if institutional design is not rigorous. In AI, the risk of deviation is particularly salient because: assessing safety, bias, robustness, and impacts depends on data and technical choices that regulators cannot always fully observe; social harms may be cumulative or systemic, extending beyond the purportedly "controlled" setting of a test; and AI's cross-sector nature increases the likelihood of fragmented oversight and overlapping mandates.

A paradox thus emerges: sandboxes can foster regulatory learning and responsible innovation when designed as supervised environments for evidence production; yet they can become "regulatory shortcuts" when they operate as prolonged zones of exception, with diffuse substantive standards, insufficient minimum transparency, and no verifiable criteria for entry, continuation, and exit. In such "shortcut" scenarios, the instrument ceases to reduce uncertainty and instead enables *compliance arbitrage* (including competitive asymmetries), externalizes risks, and erodes accountability, especially when "pilots" persist indefinitely or when participation in a sandbox functions, implicitly, as a reputational compliance seal without robust verification.

The research problem guiding this article can be stated as: how should AI regulatory sandboxes be designed and governed so that they generate regulatory learning and foster responsible innovation, without becoming a shortcut that opaquely relaxes obligations, amplifies asymmetries, and externalizes risks to the public interest?

The central hypothesis is that AI sandboxes are legitimate and useful only when structured as mechanisms of supervised progressive compliance, featuring: clear institutional design (competences, scope, and limits), auditable metrics (technical and legal-regulatory), robust governance, including multi-agency coordination where necessary, and risk-proportionate transparency, and effective exit criteria, including triggers for interruption, remediation, and, where appropriate, sanctioning. Absent these elements, sandboxes tend to operate as permanent exceptions, incentivizing indefinite pilots, weakening enforcement quality, and increasing distributive and systemic risks.

Accordingly, the object of this article is the AI regulatory sandbox as a policy instrument and public governance mechanism, examined through institutional design choices and conditions of legitimacy. The article adopts a comparative perspective, drawing on the international backdrop of "trustworthy AI" and the leading governance models currently in dispute (principle-based soft law; risk-based hard law; executive governance strategies; and more state-centered approaches), with particular attention to practical implications for metrics, transparency, coordination, and exit criteria.

The research proceeds in two complementary steps. The first consists of a critical review of the literature and doctrine on AI regulation, adaptive regulatory instruments, and sandboxes, identifying: the instrument's rationales (what it is expected to achieve), recurring criticisms (capture, opacity, arbitrage, and permanent exceptionality), and lessons and success conditions drawn from prior experiences across jurisdictions and sectors.

The second step is a comparative, design-choice–oriented examination. Rather than offering a purely descriptive country-by-country inventory, the article compares models and reference documents through an analytical matrix organized around: eligibility and scope (risk, sector, use, case or capability-based), governance and inter-institutional coordination, substantive and procedural safeguards, minimum transparency and confidentiality controls, metrics and evidentiary requirements, and entry, continuation, and exit criteria. This approach supports the extraction of comparative lessons while also illuminating recurring patterns associated with "shortcut sandboxes."

2. THEORETICAL FRAMEWORK: REGULATION, RISK, AND INSTITUTIONAL LEARNING

Regulating emerging technologies such as artificial intelligence involves a structural tension between predictability and adaptability. More prescriptive models seek to anticipate conduct and effects through detailed rules, yet they tend to become rigid in the face of accelerated technological cycles and heterogeneous use contexts; more open-ended, principle-based, or experimental approaches increase adaptability but shift the challenge to the governance of discretion, equal treatment, and control mechanisms. In AI, this tension is compounded by fragmented value chains, information asymmetries, and technical opacity, which make it costly to produce and verify evidence about safety, performance, and impacts.

The contemporary centrality of risk as the regulatory axis of AI governance does not automatically solve the problem; it rearranges it. Risk entails normative and distributive choices about which harms matter, what levels of uncertainty are tolerable, and who bears the costs of regulatory error. Moreover, AI risks often manifest not only locally but as cumulative and systemic effects, through system coupling, repeated use, network effects, institutional dependency, and externalities that fall unevenly on groups and markets. Accordingly, AI regulation must integrate technical risks (robustness, safety, bias), legal-institutional risks (accountability, due process, transparency), and systemic risks (aggregate impacts and competitive dynamics).

This is where institutional learning becomes decisive. Regulating AI is not merely about setting standards, but about building processes that generate verifiable knowledge, monitor effects, and recalibrate obligations in light of observed outcomes. Experimental instruments, such as sandboxes, are therefore justified only when they operate as environments of evidence generation and supervised progressive compliance, supported by auditable metrics, clear entry and exit criteria, and safeguards proportionate to risk. Otherwise, experimentation tends to become a permanent exception, the privatization of regulatory learning, and a regulatory shortcut that postpones duties and externalizes risks to the public interest.

2.1. Regulating emerging technologies and the limits of the prescriptive model

Regulating emerging technologies is often strained by simultaneous demands for speed, certainty, and control, which helps explain the recurring preference for prescriptive responses (detailed rules, commands, and prohibitions) when relevant risks become politically salient. In fast-moving innovation environments, however, this model faces a dilemma: the more it seeks to anticipate conduct and effects through exhaustive standards, the more it tends to produce rigidity and misfit vis-à-vis heterogeneous uses; the more it relaxes, the more it transfers the decision-making burden to concrete application and to the governance of discretion. The regulatory literature further indicates that the issue is not merely the "amount of rules," but the quality of implementation, since complex regimes can incentivize *creative compliance*, box-ticking and formalistic fulfillment that preserves the appearance of compliance without reducing the material risk that justified the intervention.

In this context, there is growing appeal to arrangements that shift part of the regulatory burden inside regulated organizations, self-regulation, meta-regulation, and corporate governance instruments, on the argument that sectoral actors hold superior technical knowledge and are therefore able to craft obligations that are more "acceptable" and, in principle, foster higher voluntary compliance. The problem is that this shift makes the model highly sensitive to organizational capacity and to the type of regulatee: sophisticated agents may exploit gaps and ambiguities, whereas less structured actors may fail due to organizational incompetence.

Moreover, the literature itself recognizes that reliance on self-regulation/meta-regulation carries recurring risks related to mandate, accountability, and procedural justice, particularly where internal structures function more as compliance rhetoric than as effective mechanisms of risk reduction[1]. Operationally, this means that less prescriptive approaches require, as a counterpart, robust verification mechanisms and enforcement design capable of dealing with different profiles of regulatees,

[1] BALDWIN, Robert; CAVE, Martin; LODGE, Martin. *Understanding Regulation*: Theory, Strategy, and Practice. 2. ed. Oxford: Oxford University Press, 2012. p. 139.

including those inclined toward amoral calculation and strategic circumvention of the spirit of norms.

In the AI domain, these limitations become more acute due to information asymmetries and the difficulty of auditing, at low cost, technical choices that affect safety, bias, robustness, and impacts. For this reason, contemporary scholarship has emphasized "adaptive" regulatory tools and a combination of instruments (soft law, auditing, certification, incentives, and oversight) to address the *pacing problem* and uncertainty[2].

Along these lines, authors argue that principles (soft law) require an "infrastructure" of credibility, internal and external mechanisms such as committees, audits, certifications, insurers, and incentives, to achieve regulatory effectiveness. At the same time, criticism of exclusively *ex post* control in AI has grown: when risks are severe or systemic, *ex post* liability tends to be slow, insufficient, and unable to prevent significant harms; accordingly, calls for *ex ante* oversight, including interventions at the stages of design, training, testing, and deployment, have gained traction [3].

Experimental instruments only respond to the limits of the prescriptive model if they are designed as mechanisms for evidence production and verifiable accountability, not as zones of exception that replace, rather than inform, substantive protection standards.

2.2. Risk-based regulation: promise, blind spots, and systemic risk

Risk-based regulation is frequently defended as the most plausible legislative technique for governing AI under conditions of uncertainty because it purports to calibrate regulatory burdens to the intensity and scope of risk, thereby preserving proportionality while avoiding blunt over-regulation. In the EU context, Ebers frames the AI Act's "risk-based approach" as a proportionality device that aims to

---

[2] CAJUEIRO, Daniel O.; CELESTINO, Victor R. R. A Comprehensive *Review of Artificial Intelligence Regulation*: Weighing Ethical Principles and Innovation. 2024. p. 35.

[3] ARBEL, Yonathan A.; TOKSON, Matthew; LIN, Albert C. *Systemic Regulation of Artificial Intelligence*. Arizona State Law Journal, v. 56, n. 2, 2024. p. 596.

balance innovation and the benefits of AI systems against the protection of fundamental values such as safety, health, and fundamental rights, while also acknowledging that key provisions of the Act may fail to operate in a "truly risk-based" manner (producing both over- and under-inclusiveness)[4].

However, risk-based schemes often embed a practical transfer of regulatory work into organizational risk management systems, effectively "delegating" parts of classification, interpretation, and compliance engineering to the regulated entity. Baldwin, Cave and Lodge note that principles-based regulation and risk-based regulation alike depend heavily on the risk management capacity of firms' internal control regimes; where that capacity is weak, compliance infrastructures may be overloaded and unable to deliver what the regulatory model assumes. They further stress that regulatees are heterogeneous: some are well-disposed but low-capacity, while others are strategically ill-disposed, including "amoral calculators" and "creative compliers" predisposed to circumvent regulatory objectives whenever advantageous[5]. This heterogeneity is not a marginal concern: it conditions whether a risk-based framework functions as a genuine governance mechanism or degenerates into formalistic, selective, or performative compliance.

A second limitation concerns the epistemic foundations of risk. In AI, risks are frequently relational (context-dependent), dynamic (evolving through updates and deployment environments), and partially latent (emerging only under scale, interaction effects, or misuse). International governance instruments therefore increasingly insist on lifecycle-based risk management, traceability, and accountability as preconditions for credible risk governance. The Organisation For Economic Co-Operation And Development (OECD)'s updated AI Principles explicitly call for traceability of datasets, processes, and decisions made during the AI lifecycle, and for a systematic risk-management approach across each phase of the AI system lifecycle on an ongoing basis, precisely to make assessment, oversight, and contestability feasible beyond

---

4 EBERS, Martin. *Truly Risk-based Regulation of Artificial Intelligence*: How to Implement the EU's AI Act. 2024. p. 1–2

5 BALDWIN, Robert; CAVE, Martin; LODGE, Martin. *Understanding Regulation*: Theory, Strategy, and Practice. 2. ed. Oxford: Oxford University Press, 2012. p. 304–305

mere declarations of compliance[6].

Finally, risk-based frameworks tend to under-specify systemic risk, i.e., the possibility that AI generates harms not reducible to discrete use cases, but rather to technology-level dynamics, diffusion, and socio-technical interdependencies. Arbel, Tokson and Lin argue that society-wide risks warrant a systemic regulatory orientation that treats AI as a technology and not solely as downstream applications[7]; they also contend that *ex post* enforcement is structurally insufficient on its own and that effective AI governance will require *ex ante* oversight and approval across the development and deployment lifecycle (architecture, objectives, training, testing, deployment), because some errors may be unfixable in hindsight.

This critique matters for the architecture of sandboxes as a sandbox can complement a risk-based regime only if it is designed as a controlled evidence-production and accountability mechanism that strengthens lifecycle oversight, rather than as a procedural exception that obscures systemic externalities.

2.3. Responsive and experimental regulation: institutional learning as a regulatory strategy

A central lesson of modern regulatory theory is that enforcement cannot be reduced to a binary choice between "persuasion" and "punishment." Rather, effective enforcement often depends on an escalatory logic, in which regulators move across a spectrum of strategies and sanctions in response to the regulatee's conduct and risk profile.

In Baldwin, Cave and Lodge's discussion of *responsive regulation*, drawing on Ayres and Braithwaite, the objective is precisely to create a "synergy" between persuasion and deterrence by structuring enforcement as a progression, often captured by the image of an enforcement pyramid, in which the intensity of regulatory re-

---

6 ORGANISATION FOR ECONOMIC CO-OPERATION AND DEVELOPMENT (OECD). *Recommendation of the Council on Artificial Intelligence* (OECD/LEGAL/0449) – 2024 Update. 2024. p. 9

7 ARBEL, Yonathan A.; TOKSON, Matthew; LIN, Albert C. *Systemic Regulation of Artificial Intelligence*. Arizona State Law Journal, v. 56, n. 2, 2024. p. 53.

sponse increases as cooperation decreases or risks rise[8]. This model is not merely tactical. It expresses an institutional claim: regulators must preserve the capacity to calibrate interventions dynamically, rather than rely exclusively on fixed *ex ante* prescriptions or uniform *ex post* sanctions.

That same institutional claim becomes even more salient under conditions of technological dynamism and epistemic uncertainty. Here, the regulatory task shifts from designing a perfect regime in advance to building systems capable of learning, revising, and improving over time. Baldwin, Cave and Lodge explicitly argue for a movement "from design to review," in which evaluation and modification are not exceptional responses to failure, but central elements of regulatory strategy.

Under an incrementalist orientation, regulators should establish a regime that appears most promising under uncertainty, measure whether it performs better than feasible alternatives, and modify regulatory strategies to effect improvement, while acknowledging the practical difficulties of benchmarking, shifting statutory mandates, and attributing responsibility in complex "mixed" regulatory regimes. This "review" paradigm provides a jurisprudential justification for regulatory instruments designed to produce evidence and feedback, thereby enabling adjustment without abandoning accountability.

Within AI governance, the case for learning-oriented regulation is reinforced by the structural limits of purely *ex post* control. Arbel, Tokson and Lin contend that *ex post* enforcement is often slow, uncertain, and weakly deterrent to prevent harms, especially where causation is difficult to trace, responsible parties are unclear, and the magnitude of harm may exceed the wrongdoer's capacity to pay. They defend systemic *ex ante* oversight across the AI lifecycle, covering architecture, objective design, training runs, testing, and deployment, precisely because some failures may become unfixable in hindsight[9]. This argument does not eliminate the need for *ex post* liability; it places it within a broader governance architecture in which prevention, verification, and iterative oversight are indispensable.

---

[8] BALDWIN, Robert; CAVE, Martin; LODGE, Martin. *Understanding Regulation*: Theory, Strategy, and Practice. 2. ed. Oxford: Oxford University Press, 2012. p. 258

[9] ARBEL, Yonathan A.; TOKSON, Matthew; LIN, Albert C. *Systemic Regulation of Artificial Intelligence*. Arizona State Law Journal, v. 56, n. 2, 2024. p. 53

Regulatory sandboxes are best understood against this theoretical background, once they are not merely "pro-innovation" policies, but potential institutional devices for structured learning under supervision. In Cajueiro and Celestino's terms, sandboxes can facilitate the identification and mitigation of risks in real-world settings before broader implementation, at least where the sandbox is embedded in a governance framework capable of generating credible evidence, preserving enforceable safeguards, and producing learning that is public-regarding rather than privately captured[10]. Accordingly, the theoretical significance of sandboxes in AI depends on whether they are designed as mechanisms of supervised progressive compliance, with clear entry and exit criteria, auditable metrics, and escalation pathways, rather than as procedural exceptions that postpone substantive duties and externalize risks.

3. INTERNATIONAL NORMATIVE BACKDROP: THE FOUNDATIONS OF "TRUSTWORTHY AI"

The consolidation of an international normative backdrop for AI governance has been shaped less by consensus on specific regulatory models and more by convergences around a shared vocabulary of "trustworthy AI." Although this convergence varies in normative density, it has clear legal relevance: it structures expectations about what should be considered acceptable in the design, deployment, and scaling of AI systems, and it operates as a reference point for national policies and international coordination. At the multilateral level, for instance, the OECD Council Recommendation on AI sets out principles applicable to all actors, including transparency and explainability, robustness, safety, and accountability, while also providing a shared vocabulary for "AI system," "AI lifecycle," and "AI actors"[11].

In parallel, the UN General Assembly Resolution of 21st March 2024 underscores the urgency of global consensus on "safe, secure and trustworthy" AI systems and encourages States and stakeholders to develop interoperable governance and safe-

---

[10] CAJUEIRO, Daniel O.; CELESTINO, Victor R. R. *A Comprehensive Review of Artificial Intelligence Regulation*: Weighing Ethical Principles and Innovation. 2024. p. 29

[11] ORGANISATION FOR ECONOMIC CO-OPERATION AND DEVELOPMENT (OECD). *Recommendation of the Council on Artificial Intelligence* (OECD/LEGAL/0449) – 2024 Update. Paris: OECD, 2024. p. 4.

guard frameworks so as to avoid regulatory fragmentation[12].

This backdrop is largely structured through principles and soft-law instruments, which do not replace but often precede and shape binding regimes. In AI, the principle-based layer tends to converge around three minimum axes: accountability, transparency (including traceability and auditability), and risk management throughout the system's lifecycle. Accountability in these documents is not confined to the final stage of use; it presupposes mechanisms that allow responsibilities to be allocated before and after development and deployment.

Scholarship systematizing the European governance landscape, for example, emphasizes that trustworthiness requires the possibility of auditing algorithms, data, and design processes, as well as mechanisms for external auditability, especially where high risks are at stake, and it explicitly links traceability to increased transparency, auditability, and explainability[13]. The OECD similarly structures its principles to require actors to promote and implement such standards "according to their roles," and its recent revision process further underscores that traceability and risk management should be treated as components of accountability, that is, as conditions for verifiable governance rather than merely declaratory commitments.

By incorporating these minimal convergences, the international debate also reinforces a methodological consequence that AI trustworthiness does not rely only on a system's status at a given moment, but on an institutional process of documentation, supervision, review, and correction. The UN Resolution itself highlights that risks may arise from improper design, development, deployment, or use, especially in the absence of safeguards, with impacts on rights, privacy, information integrity, discrimination, and accidents, thereby shifting the analysis toward preventive and continuous governance mechanisms.

For experimental instruments such as sandboxes, this implication is critical, once experimentation can be compatible with the "trustworthy AI" backdrop only if it is subordinated to minimum requirements of accountability, risk-proportionate transparency, and risk management, so as to generate evidence and responsibility rather

---

12 UNITED NATIONS. General Assembly. *Seizing the opportunities of safe, secure and trustworthy artificial intelligence systems for sustainable development*. A/78/L.49, 21 Mar. 2024. p. 4.

13 NIKOLINAKOS, Nikos. *EU Policy and Legal Framework for AI*. 2023. p. 18–21.

than operate as an opaque exception that weakens substantive standards. Principles and soft law do not "solve" sandbox design, but they delimit the bounds of what is acceptable in the name of innovation.

3.1. European Union: from "trustworthy AI" to normative densification

The European approach to the governance of artificial intelligence begins with a gradual construction of "trustworthy AI," initially articulated through *soft law* documents and later translated into binding law through the AI Act. This trajectory does not amount to a rupture, but rather to a process of normative densification.

The axes already present in the principle-based discourse, accountability, transparency, and risk management, are incorporated into positive law through procedural duties, organizational obligations, and supervisory mechanisms. The literature emphasizes that traceability of data, processes, and decisions plays an instrumental role in this design, as it connects abstract standards of trustworthiness to verifiable practices of compliance and auditability, especially in higher-risk settings[14].

The core of the AI Act is structured around risk-based regulation, which classifies AI systems according to their potential impact on fundamental rights, safety, and other relevant legal interests. This option seeks to preserve regulatory proportionality, but it faces practical difficulties. Ebers notes that reliance on predefined categories, such as the list of high-risk systems, tends to produce rigidity and lag relative to rapid technological change, and may generate intermediate zones in which systems not formally classified as "high-risk" can nevertheless produce significant effects in sensitive contexts of use. The upshot is that the model's effectiveness depends less on the taxonomy as such and more on the institutional capacity for updating, interpretation, and continuous oversight.

These tensions become more pronounced with the incorporation of general-purpose AI (GPAI) models into the European framework. Unlike systems designed for specific applications, GPAI models are built for multiple downstream uses, which makes *ex ante* risk anticipation more difficult. To address this, the AI Act introduces

[14] NIKOLINAKOS, Nikos. *EU Policy and Legal Framework for AI*. 2023. p. 21

specific obligations for GPAI providers, with a stricter regime for those classified as posing systemic risk. These obligations include the assessment and mitigation of systemic risks, continuous monitoring, reporting of serious incidents, cybersecurity measures, and corrective actions. As Ebers argues, this treatment strains the very notion of "truly risk-based regulation," because systemic risk associated with GPAI is, by definition, diffuse, emergent, and less quantifiable than risks tied to delimited applications[15].

In this context, the European Union relies on quasi–soft law instruments to operationalize hard law. The General-Purpose AI Code of Practice, developed by the AI Office, is presented as a voluntary mechanism through which providers may demonstrate compliance with the AI Act. The Code sets differentiated commitments: transparency and copyright-related obligations applicable to all GPAI providers, and additional duties concerning safety, evaluation, and incident reporting for providers of systemic-risk GPAI. The document also introduces standardized documentation templates (the *Model Documentation Form*), aiming to concentrate information relevant for supervision and to reduce information asymmetries between regulators and providers[16].

Finally, the European regime incorporates regulatory sandboxes as a supervised innovation tool, but under a rationale explicitly tied to regulatory compliance. As highlighted in the literature, the sandboxes foreseen in the AI Act are meant to enable testing and experimentation in controlled environments, especially in development and pre-market phases, with a view to supporting compliance with the Regulation and other applicable rules, enhancing regulators' supervisory capacity, and reducing legal uncertainty, particularly for small and medium-sized enterprises.

Experimentation is nonetheless conditioned on good-faith cooperation, appropriate safeguards, and the mitigation of risks to safety and fundamental rights. Thus, under the European model, the sandbox is not framed as a broad regulatory exception, but as an instrumental tool for learning and compliance, embedded in a broader

---

[15] EBERS, Martin. *Truly Risk-based Regulation of Artificial Intelligence*: How to Implement the EU's AI Act. 2024. p. 15

[16] EUROPEAN UNION. AI Office. *Third Draft of the General-Purpose AI Code of Practice*. 2025. p. 19–23.

framework of accountability and risk management.

3.2. United Kingdom: sectoral coordination, and adaptability instruments

The UK's recent model rests on a deliberate institutional choice: rather than immediately adopting a single, dense horizontal AI regime, the United Kingdom has pursued a pro-innovation and sector-led strategy, anchored in existing regulators and incremental adjustments[17]. This approach seeks to reduce regulatory friction in a fast-moving environment, preserving room for adaptation. At the same time, it shifts the system's center of gravity to coordination capacity and cross-sector coherence, since governance depends on multiple authorities operating under distinct mandates.

Within this architecture, the UK's reference framework consolidates a set of cross-cutting principles to be "tailored" and applied by regulators within their respective remits. Roberts identifies five principles: safety, security and robustness; appropriate transparency and explainability; fairness; accountability and governance; and contestability and redress[18]. In legal terms, this design reinforces the function of *soft law* as a bridge between values and implementation: the principles are not meant to operate as self-executing commands, but as directives guiding sectoral regulatory decisions, with accountability and contestability as baseline legitimacy conditions.

The immediate consequence is that the effectiveness of the UK model depends less on a rigid *ex ante* taxonomy (such as extensive high-risk lists) and more on mechanisms of coordination and consistency among regulators. Roberts notes that the White Paper emphasizes UK-wide territorial application and adopts a more "light-touch" posture than other early movers that have enacted specific horizontal

---

[17] GIKAY, Asress Adimi. *Risks, innovation, and adaptability in the UK's incrementalism*. 2024. p. 15

[18] ROBERTS, Huw; BABUTA, Alexander; MORLEY, Jessica; THOMAS, Christopher; TADDEO, Mariarosaria; FLORIDI, Luciano. *Artificial intelligence regulation in the United Kingdom*: a path to good governance and global leadership? Internet Policy Review, v. 12, n. 2, 2023. p. 8

AI statutes. While this flexibility has clear advantages, it also heightens the risk of asymmetries: divergent standards may emerge across sectors, increasing the likelihood of gaps and overlaps.

It is in this context that regulatory sandboxes appear as a characteristic instrument of the UK's regulatory innovation playbook. The UK has "pioneered" sandboxes as controlled environments in which organizations test innovations under regulatory supervision, often using real-world data, highlighting the FCA's experience since 2015 and its subsequent diffusion into other data-driven and AI-adjacent domains. ICO's sandbox as a venue for compliance guidance in data protection and mentions initiatives such as the government-supported LawTech sandbox, aimed at incubating AI solutions in the legal sector with a focus on compliance[19]. The AI Regulation White Paper commits to creating a multi-regulator sandbox specifically for AI, underscoring an intent to use supervised experimentation as a mechanism of coordination and learning across sectoral authorities.

This sandbox orientation is nevertheless ambivalent. On the one hand, it can function as an instrument of institutional adaptability, enabling supervised testing, evidence generation, more responsive regulatory guidance, and reduced uncertainty for innovators in a fast-moving AI environment. On the other hand, without minimum standards of transparency, metrics, and exit criteria, a sandbox may drift toward the risk this article seeks to delineate, becoming a regulatory shortcut that replaces (rather than informs) substantive obligations and that privatizes institutional learning.

3.3. United States: institutional fragmentation, and political volatility

The recent U.S. framework for AI is not organized, at least for now, around a horizontal statute comparable to the EU AI Act. Instead, it tends to operate through layered governance: (i) a rights-oriented and principle-based layer focused on public

[19] ROBERTS, Huw; BABUTA, Alexander; MORLEY, Jessica; THOMAS, Christopher; TADDEO, Mariarosaria; FLORIDI, Luciano. *Artificial intelligence regulation in the United Kingdom*: a path to good governance and global leadership? Internet Policy Review, v. 12, n. 2, 2023. p. 12.

protection in automated decision-making; (ii) an executive and administrative layer that mobilizes technical standards, reporting requirements, and guidance for the federal government; and (iii) a strategic security layer that treats certain models and capabilities as potential risk vectors (biosecurity, cybersecurity, critical infrastructure, and related concerns). This architecture matters for the purposes of this article because, even where it is not framed as a sandbox in the strict sense, it establishes indirect mechanisms of compliance and controlled experimentation (via standards, audits, testbeds, reporting, and procurement) that may function as institutional equivalents, or, if poorly designed, as shortcuts.

Within the rights layer, the *Blueprint for an AI Bill of Rights* (2022) sets out five principles: safe and effective systems, algorithmic discrimination protections, data privacy, notice and explanation, and human alternatives, consideration, and fallback [20]. The document clarifies that it is a public framework of values and associated practices intended to guide policy and system design where automated systems affect rights, opportunities, and access to critical needs. In the "safe and effective systems" principle, for instance, it stresses pre-deployment testing, risk identification and mitigation, and ongoing monitoring, and it expressly points to audit and documentation mechanisms as practical strategies for risk mitigation.

The executive layer gained density with Executive Order 14110 (30th October 2023), which articulates a federal strategy for "safe, secure, and trustworthy AI" and, in its statement of purpose, frames AI as simultaneously promising and dangerous, warranting risk mitigation against fraud, discrimination, disinformation, and national security threats[21]. Operationally, the Order directs National Institute of Standards and Technology (NIST) to develop guidance and best practices for safe development and deployment; it also references evaluations, audits, and benchmarks and, importantly for this article, points to the creation and use of testing environments (testbeds) to support safe development and deployment. In parallel, it establishes

---

[20] UNITED STATES. The White House Office of Science and Technology Policy (OSTP). *Blueprint for an AI Bill of Rights*: Making Automated Systems Work for the American People. Washington, DC, Oct. 2022. p. 4–5.

[21] UNITED STATES. *Executive Order 14110*: Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence. Administration of Joseph R. Biden, Jr., 30 Oct. 2023. p. 8.

reporting obligations and the disclosure of *red-team testing* results for certain dual-use foundation models, tying compliance to evidence (tests performed and mitigation measures adopted).

In 2025, political oscillation becomes a structural feature of the regulatory landscape. The Presidential Action "Removing Barriers to American Leadership in Artificial Intelligence" (23rd January 2025) reorients the agenda, revokes or mandates the review of measures associated with the prior framework, and orders the development of a new *AI Action Plan*, underscoring that, in a governance environment strongly dependent on executive instruments, regulatory priorities can shift quickly[22]. This feature has direct implications for experimental mechanisms: the greater the volatility of guidance, the greater the need for clear metrics, risk-proportionate transparency, and exit rules, so that flexibility does not become a permanent exception, and so that institutional learning is not lost with each political inflection.

Finally, the U.S. case is also marked by the absence of a central coordinating regulator with a cross-cutting mandate for the economic regulation of AI. Governance is, to a significant extent, fragmented: different agencies and levels of government address slices of the problem (consumer protection, competition, health, finance, labor, the public sector, privacy, and so forth). This can enable adaptive sectoral responses, but it raises coordination costs and tends to generate divergent expectations and enforcement gaps in intersectoral zones. In federal systems, significant norm production may shift to the subnational level and regulation may become slower and more fragmented; moreover, political-institutional factors, such as ideology and lobbying, often shape priorities and governance strategies[23]. In such an environment, predominantly executive instruments amplify the sensitivity of the framework to changes in administration, increasing perceived instability and reducing predictability about the future direction of AI governance.

---

[22] UNITED STATES. The White House. *Removing Barriers to American Leadership in Artificial Intelligence*. Presidential Action, 23 jan. 2025. p. 1.

[23] SCASSA, Teresa. *Sovereignity and the Governance of Artificial Intelligence*. UCLA Law Review Discourse, v. 71, p. 214–229, 2023–2024. p. 10–11.

3.4. Asian experiences (China, South Korea, and India): state-centered models

In contrast to the Euro-American "trustworthy AI" narrative, typically anchored in broad principles and risk-based compliance, selected Asian experiences illustrate three distinct institutional strategies for AI governance: a state-centered, infrastructure-driven approach that combines compliance duties, information controls, and security assessments (China); a comprehensive statutory framework emphasizing competitiveness and national coordination, while treating facilitation tools (including sandboxes) as part of the policy toolkit (South Korea); and a development-oriented strategy that treats transparency, explainability, and bias mitigation as adoption bottlenecks, without necessarily converting the strategy into a uniform legal regime (India). The comparative relevance lies not in ranking models, but in identifying how each architecture redistributes responsibilities, selects its regulatory objects (models, services, use-cases, sectors), and defines what counts as "safety" and "trustworthiness."

In the Chinese case, AI regulation is often described as a normative mosaic under central leadership, with a prominent role for the *Cyberspace Administration of China (CAC)*, in which sectoral rules for particular technologies (e.g., algorithmic recommendation and *deep synthesis*) operate as risk-management mechanisms through procedural and technical duties.

Among these duties, Zhang highlights *filing* requirements for certain providers of algorithmic recommendation services, including the submission of information about the algorithm and a self-assessment report, alongside obligations relating to data security and personal information protection, combined with transparency-related requirements and technical security measures[24]. The framework also incorporates security-assessment logics for certain sensitive functionalities or services, bringing regulatory control closer to an expanded notion of security governance (national, social, and informational). In this setting, experimentation is more likely

[24] ZHANG, Laney. *China*: Regulation of Artificial Intelligence. Washington, D.C.: Law Library of Congress, 2023. p. 4

to be conditioned by *ex ante* gatekeeping and structured controls than to function as a broad flexibility device.

South Korea, in turn, has moved toward a general framework statute (AI Basic Act) intended to reconcile industrial policy and trust foundations, with commitments to institutional coordination ("control tower") and safety infrastructure[25]. In this context, promotional instruments, including references to temporary permissions and regulatory sandboxes for convergent products and services, appear as part of the broader repertoire for reducing innovation friction while maintaining a baseline of obligations and standards. The comparison is instructive because the key question is not whether a sandbox exists, but how it is integrated, as a supervised exception, as a facilitation tool within industrial policy, or as a potential escape valve from general rules.

India offers an example of a national AI strategy oriented toward inclusion and development ("AI for All"), where emphasis is placed on capacity-building and adoption in priority sectors. The strategy gives particular salience to the "black box phenomenon" and treats transparency/explainability as a condition of legitimacy and public trust in the expansion of AI use in sensitive decisions[26]. This configuration tends to yield a type of governance in which "trustworthiness" is framed as a public-policy and institutional-design problem (standards, practices, capabilities, incentives, ecosystem), rather than as immediate, uniform legal compliance. For experimentation tools, the implication is straightforward: controlled testing can be justified as a learning and coordination mechanism, but its legitimacy depends on clear policy objectives, safeguards, and accountability mechanisms, otherwise pilots may perpetuate asymmetries, biases, and externalities under the label of innovation.

Overall, these cases suggest that "trustworthy AI" is not a single package, but an institutional arrangement that varies according to what is controlled (models, services, sectoral uses), how control is exercised (procedural duties, technical standards, approvals/security assessments *ex ante*), and for what purposes (rights protection,

---

[25] UNITED STATES. International Trade Administration. *South Korea Artificial Intelligence* (AI) Basic Act. Washington, D.C.: U.S. Department of Commerce, 2025.

[26] NITI AAYOG. *National Strategy for Artificial Intelligence* (#AIForAll). New Delhi: Government of India, 2018. p. 86

economic development, informational/national security). In China, for example, the emphasis on *filings* and algorithmic self-assessment for recommendation services with "public opinion attributes" or "social mobilization capabilities," alongside security-assessment mechanisms in sensitive contexts, illustrates a design in which accountability and risk are operationalized through *ex ante* gatekeeping and compliance duties, with a pronounced focus on information and data security. By contrast, India's strategy explicitly diagnoses the "black box phenomenon" and frames the legitimacy of AI in sensitive domains through explainability and institutional capacity, rather than treating trust as a purely declaratory commitment.

These contrasts indicate that sandboxes perform different functions depending on the surrounding institutional ecosystem. In arrangements centered on *ex ante* controls and state gatekeeping, experimentation tends to operate as a screening and supervision instrument, leaving limited room for broad regulatory flexibility. In more development-oriented strategies, experimentation can help reduce uncertainty, provided it is accompanied by safeguards ensuring risk-proportionate transparency, verifiable metrics, and clear exit criteria. Without these elements, pilots and controlled environments may persist indefinitely, shifting risks to the public and turning regulatory learning into a private asset.

3.5. Digital sovereignty: data, infrastructure, and governance tensions

While sovereignty is traditionally invoked to justify autonomous state action "within borders," AI (like other digital technologies) operates through cross-border flows and reorganizes the state's capacity to control and steer. Scassa therefore proposes focusing less on sovereignty as exclusive territorial jurisdiction and more on sovereignty as an analytic key for technology governance, especially because the data-driven nature of AI reshapes relations of autonomy and control. In this register, *digital sovereignty* emerges as the state's power to assert control in the digital domain, increasingly influencing contemporary AI governance debates[27].

[27] SCASSA, Teresa. *Sovereignity and the Governance of Artificial Intelligence*. UCLA Law Review Discourse, v. 71, p. 214–229, 2023–2024.

The inflection point is that digital sovereignty is not merely "the right to choose one's own path" internationally; it also describes an ability to exercise power over corporations, individuals, and even other states, precisely because technological infrastructure is indispensable and largely mediated by private actors. This helps explain why the topic is intertwined with geoeconomic disputes: policymaking takes place on a global stage in which normative/regulatory dominance tends to track economic and technological dominance, and AI governance functions simultaneously as domestic rulemaking and as a claim to international power.

Robles-Carrillo reinforces that "digital sovereignty" is a contested category, as there is no single, stable concept. In some accounts, it is less a closed legal definition than a political vision aimed at reducing dependencies in a digitized world. This agenda is typically operationalized through data policies, cybersecurity, standards, and infrastructure, rather than solely through substantive commands directed at AI. The literature the author mobilizes is revealing: digital sovereignty may be understood as an umbrella that includes pillars such as computing power, control over data, and secure connectivity, material components that condition both innovation and regulatory capacity[28].

In democratic and autocratic contexts, the invocation of digital sovereignty changes function as it may serve to preserve liberal values "in digital transformation," or it may support the consolidation of state power through digital autarky, technological isolation, and the control of citizens, including via the creation of infrastructures of control (and possible manipulation). Moreover, an inflated use of the term fosters ambiguity, once digital sovereignty can operate as a "fuzzy" label and, in some European debates, as a "Trojan horse" for protectionist measures.

The practical implication for AI regulatory architecture is that trust tools (standards, audits, transparency, risk assessment) are also evaluated by their ability to preserve institutional autonomy and avoid structural dependence on private or foreign infrastructures. If experimental environments (including sandboxes) are used as shortcuts, the sovereignty issue returns indirectly. An experiment can generate public-regarding learning only if it is governed through verifiable structures

[28] ROBLES-CARRILLO, Margarita. *Sovereignty vs. Digital Sovereignty*. Journal of Digital Technologies and Law, v. 1, n. 3, p. 673–690, 2023.

(metrics, documentation, entry and exit criteria, risk-proportionate transparency, and accountability mechanisms). Without these elements, "experimentation" tends to become a reproduction of asymmetries, where learning is privately appropriated and risks are socialized.

As a consequence, when experimentalist environments (including sandboxes) are used in AI, the sovereignty lens adds an additional requirement: the experiment must be institutionally structured to build public capacity, not merely to facilitate private testing. At a minimum, this requires an explicit regulatory objective and a testable hypothesis; transparent criteria for entry, scope, and duration, with risk-proportionate constraints; verifiable metrics and an evidence-collection plan (including logging, documentation, and traceability) that enables audit; an accountability governance arrangement (defined roles, reporting duties, incident management, and contestation mechanisms); and clear exit rules that translate learning into standards, obligations, or enforcement decisions. Absent these elements, experimentation is likely to reproduce asymmetries: it shifts costs and risks to the public, preserves opacity, and ultimately deepens technological dependencies that weaken regulatory capacity.

4. INSTITUTIONAL DESIGN OF AI SANDBOXES: GOVERNANCE, METRICS, AND LIMITS

The adoption of sandboxes for AI is often justified as an institutional response to accelerating technological change and to the difficulty for law and regulation to "reconnect" with new social and economic dynamics at the same speed as markets. On this view, a sandbox is not mere regulatory "flexibility", it is a method for generating information and reducing uncertainty through controlled testing, allowing regulators to calibrate admissible risk levels on the basis of evidence collected throughout the experimental cycle. This perspective shifts the debate from "permit or prohibit" to "how to learn safely," aligning experimentalism with a more collaborative, continuously monitored, and iteratively adjustable mode of regulatory decision-making[29].

[29] COBRA, Thiago Guimarães de Barros. *Sandbox Regulatório*: Regulação Experimental da Saúde. Dissertação (Mestrado) — Faculdade de Direito, Universidade de São Paulo, São Paulo, 2024. p. 89-90.

For such an arrangement to produce institutional learning (rather than erode safeguards), the starting point is procedural: clearly defining how entry, operation, and exit work within the experimental environment. From a legal design standpoint, it is recommended to specify the selection (or, exceptionally, qualification) process, eligibility criteria, safeguards, risk and responsibility management, temporary authorizations, supervision and monitoring, review and adjustment mechanisms, a discontinuation plan, evaluation, and transition to definitive authorizations, supported by active and passive transparency tools.

In AI, this is particularly sensitive because risks may materialize diffusely (e.g., discrimination, decision opacity, data leakage, performance degradation in real-world contexts), requiring safeguards built around limitations on scale, operational scope, territory, and number of participants, as well as risk-management routines capable of responding to unforeseen events[30].

Governance, in turn, must be treated as the infrastructure of learning: it determines who decides, how decisions are made, and how conflicts are resolved, and how collaboration among the regulator, participants, and affected stakeholders is ensured, with transparency and accountability compatible with the exceptional nature of the regime. This governance may be sectoral or multi-sectoral, permanent (through successive cohorts) or temporary (through calls for participation), and frequently relies on deliberative and operational committees to separate strategic decisions, technical oversight, and day-to-day execution.

Brazilian experience has already highlighted the importance of minimum procedural standards (stages, objective entry criteria, implementation, monitoring, and exit) as a condition to prevent sandboxes from becoming mere regulatory weakening through governance and transparency failures. This is precisely where the "shortcut" risk arises: without transparency regarding design, collected data, and outcomes, including "less significant" findings, the sandbox loses legitimacy, reduces public scrutiny, and may normalize exceptions without an empirical basis.

The metrics-and-evaluation axis closes the cycle and delineates the boundary

---

[30] BRAZIL. Attorney General's Office (AGU). *Regulatory Sandbox Reference* Guide. Brasília, DF: Labori/AGU; Ministry of Development, Industry, Trade and Services (MDIC), 2024, p. 45-46

between responsible experimentation and permissiveness. Evaluation is not a rhetorical step, it determines feasibility, safety, and impacts, enabling (or not) the transition from testing to broader market deployment. It is recommended to develop operational indicators (efficiency, quality, costs, response times), regulatory indicators (effects of flexibilities on markets, consumers, and competition), and social/environmental indicators (impacts on vulnerable groups, employment, externalities), accompanied by data standardization and interim and final reports that document the entire experimental cycle.

The institutional value of a sandbox relies on translating these outcomes into regulatory policy. The experiment should feed a reusable knowledge base (with appropriate safeguards for confidentiality, data protection, and intellectual property), highlighting successes and failures to guide future cohorts and reduce information asymmetries[31].

4.1. Purposes of AI sandboxes: regulatory learning and supervised compliance

An AI sandbox is justified when its primary purpose is to generate evidence under controlled conditions, reducing uncertainty for both regulators and market actors. The experimental cycle should make it possible to test, at limited scale, which risks actually materialize, which safeguards are effective, and which regulatory requirements are necessary (or dispensable) for a given use context. On this view, a sandbox is not synonymous with flexibility; it is an information-generation method oriented toward regulatory decision-making, including the decision not to authorize.

In addition, a sandbox can operate as a mechanism of supervised progressive compliance. In AI, relevant duties (documentation, monitoring, incident response) are implemented and refined over time; the experimental environment allows regulators to require a minimum core of obligations from the outset, while permitting controlled adjustments during the cycle, provided that criteria, records, and supervi-

[31] SOARES, Evandro da Silva. *O sandbox como ambiente experimental de tecnologias emergentes, adaptação regulatória e inovação.* Tese (Doutorado) — Universidade Presbiteriana Mackenzie, São Paulo, 2025. p. 9-11.

sion are in place. This structure also helps reduce information asymmetries, organizing reporting and documentation so that private technical knowledge becomes verifiable regulatory evidence[32].

Finally, the arrangement fulfills these purposes only if the cycle results in reusable institutional learning. This requires recording findings (including negative ones), consolidating evidence in reports, and translating results into standards, guidance, or requirements that can be applied beyond the pilot, with conversion of outcomes into regulatory policy and an accessible knowledge base.

4.2. Eligibility and selection: scope boundaries and risk-proportionate design

The eligibility and selection stage is where the sandbox stops being a generic idea and becomes a governable arrangement. The central task is to align risk proportionality with actual supervisory capacity. In AI, it is not enough to prioritize what is merely innovative, but rather what can be tested with adequate control, clear objectives, and manageable risk throughout the cycle. For that reason, entry criteria should reflect both the prospective public value and the internal quality of the proposed experiment (data handling, monitoring, mitigation measures, and responsibility allocation).

Selection also requires subjective limits, defining how many participants will be admitted and what minimum capabilities they must demonstrate to comply with reporting, audit, and incident-response obligations[33]. Constraining the participant pool is not a bureaucratic detail; it is a condition for effective oversight and for reducing unequal treatment. In AI, this can be translated into baseline requirements for internal governance and operational capacity (version control, data controls, accountable teams, and readiness to implement corrective actions), as well as evidence

---

[32] BRAZIL. Attorney General's Office (AGU). *Regulatory Sandbox Reference Guide*. Brasília, DF: Labori/AGU; Ministry of Development, Industry, Trade and Services (MDIC), 2024, p.54.

[33] COBRA, Thiago Guimarães de Barros. *Sandbox regulatório*: regulação experimental da saúde. 2024. Dissertação (Mestrado em Direito do Estado) – Faculdade de Direito, Universidade de São Paulo, São Paulo, 2024, p. 97–98.

that the project fits the cohort's objectives and scope.

In parallel, objective limits are indispensable, as what exactly is being tested, in which context, with what data, and affecting which populations. The purpose is to prevent the pilot from expanding "by drift" into unassessed purposes, especially when uses affect rights or involve vulnerable groups. Accordingly, the scope should delimit purpose, enabled functions, data flows, degree of autonomy, and scale limits (number of users/decisions), establishing a clear perimeter for tailoring safeguards and for producing comparable results.

The design should include minimum entry conditions and clear grounds for rejection and early termination. This requires, from the outset, a basic execution and risk-management plan: eligibility criteria, safeguards, responsibilities, supervision, and evaluation mechanisms must be set out in the sandbox's normative instrument, so that selection operates as an effective filter against projects seeking only a "regulatory window" without an evidence-based counterpart.

4.3. Safeguards and risk management throughout the experimental cycle

In AI, safeguards should be treated as an integrated package of risk containment, not as a formal annex to a call for participation. Sandbox limits (scale, duration, functional scope, affected population) are the first line of defense; yet they are insufficient where risks materialize diffusely or cumulatively (e.g., discrimination, opaque decisions, data leakage, performance degradation in real-world settings). For that reason, the cycle must combine structural limits with active risk-management routines capable of detecting, mitigating, and, where necessary, interrupting the experiment[34].

A second safeguard axis concerns material accountability for harm. If the experiment exposes users/consumers to additional risk, the arrangement should provide proportionate guarantees: a clear duty to repair, compensation mechanisms, and, where appropriate, financial guarantees (e.g., insurance, escrow, bond, a compensa-

[34] BRAZIL. Attorney General's Office (AGU). *Regulatory Sandbox Reference Guide*. Brasília, DF: Labori/AGU; Ministry of Development, Industry, Trade and Services (MDIC), 2024, p.45-46.

tion fund, or equivalent coverage) to ensure compensation for harms caused during the pilot. The function of these guarantees is not only compensatory; it also disciplines incentives, preventing the costs of risk from being shifted onto participants and reinforcing the seriousness of mitigation commitments.

In the UK FCA Regulatory Sandbox[35], consumer protection is operationalized through bespoke safeguards negotiated for each test, explicitly aimed at mitigating potential harm during and after testing. The FCA reports concrete examples from early cohorts such as additional capital requirements, systems penetration testing, and secondary review of robo-advice by a qualified adviser, illustrating that sandbox flexibility is paired with tailored risk controls rather than assumed to be self-limiting.

The third axis is incident management, through contingency planning and reversal mechanisms. In AI, this typically includes objective triggers to pause or terminate testing, incident-response procedures (containment, remediation, communication), root-cause analysis, and continuity measures (human fallback, non-automated alternatives, return to the prior workflow). In practical terms, the safeguard depends on operational readiness: it is not enough to have a plan; participants must be able to demonstrate that teams, governance, and resources exist to execute it under pressure, with audit trails and defined timelines.

The most sensitive safeguard is full information on risk exposure for users/consumers participating in the experiment. In an experimental environment, transparency is not optional, it is a condition of legitimacy. This requires clear and verifiable communication of the experimental nature of the service, the scope of testing, the main risks and limitations, mitigation measures, complaint and contestation channels, and the conditions for ending or changing the experiment. It is also advisable to specify what will be disclosed publicly (aggregated results, safety findings, corrective measures) and what may be restricted on confidentiality grounds, with justification.

In digital health sandboxes, legitimacy is tied to explicit risk governance rather than to "innovation framing". The World Bank's implementation brief stresses that

---

35 FINANCIAL CONDUCT AUTHORITY (FCA). *Regulatory sandbox*: lessons learned report. London: FCA, 2017, p. 3.

sandbox operators should define a risk model for implementation, anticipate unintended consequences, and establish contingency and mitigation mechanisms aimed at protecting all stakeholders, especially patients and the public. It further notes that, even with mitigation in place, patients and the public must understand and be conscious of the risks associated with participating in sandbox testing, linking safety, transparency, and trust as practical conditions of acceptability[36].

4.4. Governance, continuous oversight, and incident documentation

The governance of an AI sandbox should be conceived as a continuous decision-making arrangement, rather than as sporadic oversight of selected projects. This requires institutional structures that accompany the experiment from design to exit, with the authority to adjust safeguards, impose corrective actions, pause activities, and consolidate evidence. Good-practice guidance on sandboxes emphasizes organizing governance by functions (strategic deliberation, operational monitoring, and evaluation), separating decision, execution, and control so the cycle does not depend on ad hoc choices[37].

A robust institutional route is the establishment of steering committees within the regulator. These committees should not exist merely to "approve projects," but to design the experiment, set entry criteria, delimit scope, impose limits on scale and duration, define safeguards, and establish the evidence plan (metrics, data, and reporting format). Their composition should reflect AI's multi-dimensional profile, incorporating expertise in information security/cybersecurity, economics and competition, social impacts (including bias and discrimination), and law and public governance. This structure reduces the risk of one-dimensional decisions and strengthens proportionality judgments throughout the cycle, without transferring the regulator's mandate to external actors.

Governance should also explicitly include channels for protecting participating

---

[36] WORLD BANK. *Regulatory sandboxes for digital health*. Washington, DC: World Bank, 2025. p. 9

[37] BRAZIL. Attorney General's Office (AGU). *Regulatory Sandbox Reference Guide*. Brasília, DF: Labori/AGU; Ministry of Development, Industry, Trade and Services (MDIC), 2024, p.36.

users/consumers. Where experiments expose consumers to additional risks, it is advisable to structure sandbox-specific complaint and incident-handling procedures, as a dedicated channel, response deadlines, traceable records, escalation criteria (from the participant to the regulator), and remediation mechanisms. A sandbox-specific complaint function serves a dual purpose, it protects users and produces regulatory evidence about failures, asymmetries, and externalities that would otherwise remain underreported or invisible.

On the regulated side, accountability becomes concrete through ongoing reporting duties. Beyond general requirements of risk-proportionate transparency, participants should be required to report agreed indicators, material changes in the model/system (versions, data, operational parameters), adverse events, and corrective measures. In practice, this implies periodic reports (monthly or tied to defined milestones), not only a final report. The final report consolidates findings, but interim reporting enables dynamic correction and gives the regulator real supervisory capacity, reducing the risk that learning becomes concentrated only at the end of the cycle.

Sandbox governance should incorporate a regime of calibrated disclosure, publishing the design, objectives, safeguards, and aggregated results, restricting information only where strictly justified (e.g., security, personal data, intellectual property). The emphasis on broad disclosure, and on having explicit justification for any restrictions, matters because it supports public legitimacy and prevents sandbox exceptionalism from becoming institutional opacity[38].

4.5. Exit rules and institutionalizing learning: from testing to decision

Exit rules are the element that prevents a sandbox from turning into a parallel regime of regulatory flexibility. An experimental programme preserves legitimacy only when it is designed with a finite duration, objective closure conditions, and a clear menu of end-of-cycle decisions. A recurrent weakness lies precisely in the final stage: pilots that are extended by inertia, renewals without criteria, and participants kept in an exceptional status for an indefinite period, producing asymmetries and

[38] BRAZIL. Attorney General's Office (AGU). *Regulatory Sandbox Reference Guide*. Brasília, DF: Labori/AGU; Ministry of Development, Industry, Trade and Services (MDIC), 2024, p.31.

shifting risks onto users and competitors without a proportionate return in control and evidence[39].

For that reason, exit rules must be defined from the outset along three minimum dimensions. The first is temporal: the overall duration, interim milestones, and the possibility of extension only under explicit criteria and technical justification. The second is material: triggers for early termination (serious incidents, breach of safeguards, technical infeasibility, or unmanageable risk), coupled with remediation duties and safe rollback mechanisms. The third is decisional: the catalogue of possible outcomes should be known in advance, for example, definitive authorization (with or without conditions), an exceptional extension with adjusted scope, closure with return to the general regime, or prohibition of the tested use where risks and benefits prove incompatible.

The exit decision, however, is not merely a "final act", it must be evidence-driven throughout the cycle. This requires anchoring exit rules in minimum evaluative criteria, what evidence is needed to justify scaling, what failures require adjustment, and what risks mandate closure. In AI sandboxes, the literature emphasizes that transitioning successful innovations into the ordinary regime depends on transparent evaluation criteria, and empirical experience points to stage-based evaluative instruments as a way to reduce arbitrariness and enable comparisons across cohorts and programmes[40]. This is where the next chapter connects: metrics do not "replace" regulatory judgment, but provide the evidentiary basis that prevents exit decisions from being captured by success narratives or by pressures for continuity.

A further critical point is the post-sandbox transition. Where results are positive, moving from pilot to the ordinary regime should be conditioned on compliance requirements and post-deployment monitoring, avoiding a leap in scale without proportionate safeguards. Where results are negative or ambiguous, exit should provide for an orderly discontinuation, as communication to users, return to alterna-

---

[39] COBRA, Thiago Guimarães de Barros. *Sandbox regulatório*: regulação experimental da saúde. 2024. Dissertação (Mestrado em Direito do Estado) – Faculdade de Direito, Universidade de São Paulo, São Paulo, 2024, p. 96.

[40] MAZO FORERO, Juan. *Sandboxes regulatorios en inteligencia artificial*. Bogotá: Universidad Externado de Colombia, Centro de Estudios sobre Genética y Derecho, 2025, p.2.

tive workflows, preservation of records for audit and *ex post* evaluation, and reparations where harm occurred[41]. Identifying closure as a structural weakness reinforces that the final phase must be treated as part of sandbox institutional design, not as a residual afterthought.

Empirically, exit is structured as a distinct phase rather than an informal "end of testing." Evidence on the FCA process describes regular reporting requirements during testing, followed by submission of a final testing report and exit, after which firms may apply for permanent authorization. This sequence is an example of how sandbox cycles can be designed to convert experimentation into a verifiable decision point, rather than extending by inertia[42].

Exit rules also function as a mechanism of equality and competitive integrity. A sandbox that does not close (or that closes selectively) creates undue advantages, encourages regulatory arbitrage, and distorts third parties' regulatory expectations. By tying timelines, closure conditions, and decision options to verifiable evidence, the regulator reduces the risk of a "regulatory shortcut" and reinforces the premise that sandbox exceptionalism must generate public learning and replicable decisions, rather than indefinite temporary permissions.

5. METRICS, EVIDENCE, AND THE EVALUATION OF EXPERIMENTALISM IN AI

In AI experimentalism, assessing results is not merely about confirming that a system "worked." The core task is to generate verifiable evidence of how the system behaved in real-world conditions, which risks materialized (or were prevented), and what effects fell on users, markets, and third parties. Without baseline metrics and documentation, the cycle tends to produce only a temporary permission that is difficult to audit and of limited value for decisions on continuation, adjustment, closure, or scaling.

---

41 BRAZIL. Attorney General's Office (AGU). *Regulatory Sandbox Reference Guide*. Brasília, DF: Labori/AGU; Ministry of Development, Industry, Trade and Services (MDIC), 2024, p.54.

42 CORNELLI, Gabriele; FROST, Jon; GAMBACORTA, Leonardo; RAUCH, Michael; SONG, Zhongxia. *Regulatory sandboxes and fintech funding*: evidence from the UK. Basel: Bank for International Settlements, 2020. (BIS Working Papers, n. 901), p. 10.

Evaluation should therefore combine technical indicators (performance and stability), security indicators (incidents and corrective actions), and impact indicators (benefits, costs, and externalities), always with baselines and comparison criteria. In AI, updates and adjustments during the cycle can materially change the risk profile; accordingly, version traceability, incident logs, and periodic reports are part of the experimental method itself, not a post-hoc formality.

It should be noted, however, that the precision of these metrics depends directly on the quality of the data collected, which in turn depends on the experimental design: which variables are collected, at what granularity, with what frequency, and under which integrity controls. For this reason, the evidence plan should be jointly built by regulator and regulated entity through the sandbox's steering committee, which defines the perimeter of testing, reporting duties, and the documentation required for results to be comparable and auditable.

Institutional value relies on translating findings, including failures and negative results, into reusable regulatory knowledge, with risk-proportionate transparency and clear justifications for any confidentiality constraints. Metrics and documentation are what allow experimentation to be converted into replicable decisions rather than prolonged exceptions.

5.1. Performance, robustness, and operational security: minimum evidence for a reliable AI test

The evaluation of an AI experiment must demonstrate, first and foremost, how the system behaves under real-world conditions, not merely whether it achieves accuracy in a controlled setting. This requires evidence of operational performance (service quality, response times, efficiency, and stability), recorded consistently so as to enable comparison throughout the cycle and across projects. Good-practice guidance recommends that regulators require a minimum set of operational indicators, complemented by regulatory indicators (effects of any flexibilities) and social/environmental indicators, precisely to avoid impressionistic assessments and to support evidence-based decisions on continuation, adjustment, closure[43].

[43] BRAZIL. Attorney General's Office (AGU). *Regulatory Sandbox Reference Guide*. Brasília, DF:

In AI, however, this technical assessment is inseparable from robustness: performance degradation over time, sensitivity to changes in data and operating context (*drift*), and the effects of incremental updates to the model/system. For this to be measurable, evidence depends on traceability (versions, changes, relevant parameters) and on the integrity of data collection. Accordingly, experimental design should specify from the outset what will be collected, how it will be computed, and how it will be reported, ideally with standardized formats, methods, and reporting templates that enable comparability and auditability.

A third axis concerns operational security and incident-response capability. Good practice points to the need for continuous and detailed documentation, including interim reports in addition to the final report, so that relevant occurrences (failures, incidents, user complaints) do not become concentrated only at the end of the cycle. Moreover, sandbox effectiveness depends on clear risk-management mechanisms supported by specialized human and technical resources, an especially salient condition in AI, where typical risks (privacy, operational transparency, and fairness) can materialize diffusely and cumulatively during use.

Finally, the measurement of outcomes should also be read as a proxy for the quality of the experimental arrangement itself[44]. Brazilian experience with sandbox maturity assessment shows that adherence to good practices can be tested through structured instruments (e.g., criteria grouped by phases/steps of the cycle, including implementation/monitoring, evaluation, and regulatory decision-making), allowing gaps to be identified and discretion to be reduced, while also indicating that cycles tend to fail where reporting, monitoring, and evaluation are not sufficiently standardized[45].

Labori/AGU; Ministry of Development, Industry, Trade and Services (MDIC), 2024, p.53.

44 MAZO FORERO, Juan. *Sandboxes regulatorios en inteligencia artificial*. Bogotá: Universidad Externado de Colombia, Centro de Estudios sobre Genética y Derecho, 2025, p.3.

45 PORTELLA, Isabella Silva Di Jorge. *Sandbox regulatório no Brasil*: análise da implementação, dos modelos e da maturidade nas agências reguladoras federais. 2025. Tese (Doutorado Profissional em Políticas Públicas) –Escola Nacional de Administração Pública (Enap), Brasília, 2025, p. 95.

5.2. Socioeconomic impacts and distributive effects: benefits, costs, and externalities of the experiment

Evaluating an AI experiment cannot be limited to technical performance, it is necessary to assess the test's external effects, benefits, costs, and risks borne by users, markets, and third parties. For that purpose, it is advisable to define baselines and comparison criteria such as before/after, with/without, so that pilot success is not defined solely by the provider's internal metrics. Good sandbox practice recommends using operational, regulatory, and social indicators, with documentation and reporting throughout the cycle, to support evidence-based decisions on continuation, adjustment, closure, or scaling.

From a socioeconomic perspective, a first set of results should capture public value and efficiency: service quality, access, reduced costs and time, and process improvements. Those benefits, however, must be weighed against costs and externalities: frictions for contesting decisions, shifts of compliance and monitoring costs, technological dependence, and competitive effects (e.g., lock-in and entry barriers). In AI, material impacts often do not appear in averages: gains may coexist with localized quality deterioration or increased social costs for certain segments.

Accordingly, assessment should be sensitive to distributive effects and to risks of social distortion: bias, indirect discrimination, profiling, and segmentation that amplifies asymmetries. Practically, this means monitoring how benefits and errors are distributed across user profiles and use contexts, whether deployment creates "zones of opacity" or constrains the exercise of rights, and whether certain groups are exposed to systematically worse decision or treatment patterns. The literature on AI sandboxes stresses that privacy, transparency, and rights-protection issues can accumulate during use, requiring robust institutional design and transparent evaluation criteria.

5.3. Evidence governance and cycle integrity: reporting, auditability, and comparability of results

Experimental AI programmes require an evidence architecture that makes out-

comes verifiable, comparable, and decision-ready. Because AI systems may change during the cycle, through model updates, data shifts, or operational adjustments, evaluation depends on traceability and standardized records capable of showing what changed, when, and with what effects. In this setting, evidence governance is not ancillary: it is the condition that allows results to sustain regulatory decisions on continuation, adjustment, closure, or scaling.

A core requirement is a structured reporting and documentation regime. This includes periodic reporting (by milestones and/or fixed frequency), version documentation, incident logs, and records of corrective actions, all subject to minimum content and format standards. Good-practice guidance stresses interim and final reporting, together with data standardization, precisely to ensure that feasibility, safety, and impact assessments are grounded in consistent evidence rather than impressionistic narratives[46].

Auditability depends on data integrity controls and on the ability to reconstruct relevant events. Collection protocols, validation mechanisms, and traceability of changes must allow reviewers to identify causal pathways behind outcomes, especially where risks materialize cumulatively or indirectly. In AI, this reconstruction function is essential to avoid attributing success to aggregate performance while overlooking degradation, subgroup harms, or incident patterns that only appear in longitudinal records[47].

Institutional learning requires that findings, positive and negative, be organized for reusability. Converting results into guidance, conditions of use, and future supervisory practices presupposes documentation that supports cross-case comparison. Stage-based evaluation instruments, with explicit criteria and scoring scales, have been proposed as a means of measuring adherence to good practices and reducing discretionary variance across programmes, enabling systematic comparison and

---

[46] BRAZIL. Attorney General's Office (AGU). *Regulatory Sandbox Reference Guide*. Brasília, DF: Labori/AGU; Ministry of Development, Industry, Trade and Services (MDIC), 2024, p.54.

[47] COBRA, Thiago Guimarães de Barros. *Sandbox regulatório*: regulação experimental da saúde. 2024. Dissertação (Mestrado em Direito do Estado) – Faculdade de Direito, Universidade de São Paulo, São Paulo, 2024, p. 90.

clearer identification of gaps in implementation and monitoring[48].

Evidence governance also protects equality and integrity in experimental regimes. Where reporting is inconsistent or documentation is incomplete, exit decisions become opaque and susceptible to selective continuation, undermining competitive neutrality and incentivizing arbitrage. By contrast, standardized reporting, audit-ready records, and reusable documentation reduce asymmetries and help ensure that sandbox exceptionalism yields public learning rather than prolonged permissions.

6. WHEN THE SANDBOX BECOMES A DANGEROUS "SHORTCUT": RISKS AND CONDITIONS OF LEGITIMACY

The legitimacy of the sandbox rests on a simple premise: exceptional treatment is justified only if it generates verifiable public learning. Where institutional design is consistent, combining safeguards, ongoing oversight, and documentation, experimentation can reduce uncertainty and improve the quality of regulatory decision-making. Without these conditions, controlled flexibility may degrade into informal authorization, with asymmetric relaxation of controls and risk-shifting.

The main driver of this degradation is the fragility of the cycle at its key decision points: evaluation and closure. Accumulated experience with sandbox programmes shows that, where *ex ante* criteria and standardized evidence are lacking, pilots tend to be prolonged by inertia or renewed without clear parameters, generating competitive asymmetries and decisional opacity[49]. In AI, this risk is compounded by externalities that materialize in use, bias, profiling, incremental degradation, and distributive effects, which require traceability and accountability to avoid remaining invisible to the decision-making process[50].

---

[48] PORTELLA, Isabella Silva Di Jorge. *Sandbox regulatório no Brasil*: análise da implementação, dos modelos e da maturidade nas agências reguladoras federais. 2025. Tese (Doutorado Profissional em Políticas Públicas) –Escola Nacional de Administração Pública (Enap), Brasília, 2025, p. 90.

[49] COBRA, Thiago Guimarães de Barros. *Sandbox regulatório*: regulação experimental da saúde. 2024. Dissertação (Mestrado em Direito do Estado) – Faculdade de Direito, Universidade de São Paulo, São Paulo, 2024, p. 97.

[50] BRAZIL. Attorney General's Office (AGU). *Regulatory Sandbox Reference Guide*. Brasília, DF:

In this respect, the debate shifts from the sandbox's institutional engineering to its conditions of legitimacy: when the exception remains tied to evidence and control, and when it begins to operate as a shortcut. AI sandboxes are defensible only as procedures for producing evidence and audit-ready decisions, capable of closing cycles on the basis of verifiable criteria and preventing experimentalism from turning into permanent exceptionalism[51].

6.1. Exceptionalism becomes the rule: cycle perpetuation and erosion of exit rules

The first form of sandbox degradation occurs when the exception is prolonged and begins to operate as a practical rule. Instead of a finite testing procedure, the programme stabilizes as a "permanent provisional regime," with successive renewals and without verifiable criteria for transition into the ordinary regime or for discontinuation. In this scenario, exceptional treatment ceases to be a learning method and becomes a continuous channel for regulatory flexibility, undermining the instrument's justification.

This drift is typically enabled by weak *exit rules*: extensions by inertia or open-ended drafting, lack of interim milestones, and the absence of defined regulatory consequences at the end of the cycle. Empirical experience identifies closure as one of the most fragile points in sandbox programmes, requiring explicit discontinuation plans and clear decision rules; and good-practice guidance recommends that the normative instrument specify supervision, evaluation, and transition (or closure) mechanisms consistent with temporary authorizations[52].

In AI, perpetuation is especially problematic because material risks often emerge in use and over time, so extending cycles without standardized evidence and trans-

Labori/AGU; Ministry of Development, Industry, Trade and Services (MDIC), 2024, p.53.

51 PORTELLA, Isabella Silva Di Jorge. *Sandbox regulatório no Brasil*: análise da implementação, dos modelos e da maturidade nas agências reguladoras federais. 2025. Tese (Doutorado Profissional em Políticas Públicas) –Escola Nacional de Administração Pública (Enap), Brasília, 2025, p. 84.

52 BRAZIL. Attorney General's Office (AGU). *Regulatory Sandbox Reference Guide*. Brasília, DF: Labori/AGU; Ministry of Development, Industry, Trade and Services (MDIC), 2024, p.53.

parent evaluation criteria increases user exposure to externalities and amplifies competitive asymmetries. Containing the "shortcut" therefore requires tying closure to auditable evidence as effective timelines, termination triggers, and consistent reporting capable of supporting decisions based on feasibility, safety, and impacts.

6.2. Competitive asymmetry and "sandbox shopping": regulatory arbitrage and market distortions

A second legitimacy risk arises when the sandbox begins to operate as a competitive advantage mechanism rather than as an evidence-producing procedure. By granting flexibilities (even if temporary), the programme may reduce compliance costs, accelerate time-to-market, and enable early scaling under exceptional conditions. If such advantages are not strictly conditioned on safeguards, reporting duties, and exit criteria, they tend to generate competitive asymmetry: participants operate under a less burdensome regime while competitors remain bound by ordinary requirements, with relevant effects on prices, innovation, and barriers to entry.

This scenario worsens when cycle continuity is uncertain or selective. Extensions by inertia and weak closure criteria magnify the participant's competitive gain and heighten the risk of agenda capture: the debate shifts toward maintaining the exceptional regime rather than evaluating the experiment. For that reason, good-practice guidance insists that normative instruments address temporary authorizations, supervision, evaluation, and discontinuation/transition, precisely to prevent exceptionalism from turning into a "prolonged license" with structural market effects.

In AI, an additional concern is "sandbox shopping", a form of regulatory arbitrage, the strategic search for programmes (or jurisdictions) with lighter requirements, lower transparency, or weaker oversight capacity. The literature on AI sandboxes warns about regulatory arbitrage and emphasizes the need for transparent evaluation criteria and robust institutional design, including coordination mechanisms that reduce incentives for opportunistic migration across experimental regimes[53].

[53] HANDRLICA, Jakub; SHARP, Vladimír; NEŠPOR, Jan. *Forum shopping in regulatory sandboxes and the perils of experimental law-making*. Juridical Tribune – Tribuna Juridica, v. 13, n. 3, p. 408–

Empirical and policy-oriented literature has already flagged "sandbox shopping" as a predictable by-product of fragmented sandbox ecosystems. Truby warns that leaving AI sandbox implementation to separate national authorities can incentivize actors to exploit lax sandbox environments, thereby encouraging "sandbox shopping" and undermining uniformity of protection. A related legal analysis frames the same dynamic as forum shopping: firms may strategically seek the most favorable sandbox entry criteria and supervisory intensity, which increases regulatory arbitrage risks unless authorities coordinate baseline conditions[54].

6.3. Opacity and capture: informational asymmetry and the weakening of public learning

The specific risk here is not low transparency in abstract, but the sandbox's structural informational asymmetry. Regulatee controls most of the data, metrics, and evidence about system behavior and impacts, while the regulator tends to decide on the basis of reports produced by the participant itself. Without counterweights, this architecture turns experimentation into a black box and shifts learning into the private domain, reducing scrutiny and expanding the space for weakly verifiable decisions on continuation, adjustments, and scaling[55].

This asymmetry is a typical vector of capture in experimental settings, not necessarily corruption, but cognitive and operational dependency of the regulator on the participant, reinforced by informal channels, low standardization of reporting, and the absence of trails that allow reconstruction of relevant events. For that reason, good-practice guidance insists on governance arrangements that separate decisional and operational functions, record relevant deliberations, and impose minimum documentation and reporting standards, precisely to reduce the likelihood that cycle management becomes driven by pressure for results.

425, 2023.

54 TRUBY, Jon; BROWN, Rafael; IBRAHIM, Imad Antoine; CAUDEVILLA, Oriol. *A sandbox approach to regulating high-risk artificial intelligence applications*. SSRN, 12 nov. 2021. p. 36.

55 BRAZIL. Attorney General's Office (AGU). *Regulatory Sandbox Reference Guide*. Brasília, DF: Labori/AGU; Ministry of Development, Industry, Trade and Services (MDIC), 2024, p.54.

In AI, the risk is compounded because material impacts (bias, profiling, incremental degradation, and distributive effects) may not appear in aggregate metrics and can therefore remain invisible unless evidence is structured to reveal them. The literature on AI sandboxes stresses the need for robust institutional design and transparent evaluation criteria to mitigate risks linked to rights, privacy, and operational transparency during use. As an institutional response, stage-based evaluation instruments (with explicit criteria and scoring scales) help reduce discretionary variance and require a common minimum of comparable evidence across cycles and participants, preventing sandboxes from producing mere temporary permissions without public learning.

CONCLUSION

This article began from a central problem: defining when regulatory sandboxes applied to AI remain legitimate as instruments of institutional learning, and when they degrade into shortcuts, exceptions that relax controls without generating sufficient public evidence to justify continuation, transition, or scaling. The working hypothesis was that this boundary does not depend on the sandbox label itself, but on the design of the decision cycle: exceptional treatment is defensible only if it is tied to safeguards, documentation, and verifiable closure criteria capable of converting experience into an audit-ready decision.

The decisive point is not the existence of a controlled environment, but the capacity to close the cycle on the basis of criteria, preventing extensions by inertia that turn exception into practice and generate asymmetries. In AI, this risk is structurally higher because salient effects materialize in use and over time, including bias, profiling, and distributive impacts, so regulatory decisions tend to be captured by performance narratives unless evidence governance is in place. Metrics are not a technical add-on; they are the condition for comparable evaluation and for an evidence-based exit decision.

AI sandboxes are defensible only as procedures for producing evidence and audit-ready decisions, not as spaces of regulatory tolerance. This requires treating as the methodological core: data and documentation integrity, traceability of changes and occurrences, calibrated and reasoned transparency, and effective user/consumer

protection, with adequate information about risk exposure and remediation/redress mechanisms proportionate to risk. Absent this bundle, exceptionalism tends to socialize harms, privatize learning, and generate undue competitive advantages.

REFERENCES

ARBEL, Yonathan A.; TOKSON, Matthew; LIN, Albert C. *Systemic Regulation of Artificial Intelligence.* Arizona State Law Journal, v. 56, n. 2, 2024.

BALDWIN, Robert; CAVE, Martin; LODGE, Martin. *Understanding Regulation*: Theory, Strategy, and Practice. 2. ed. Oxford: Oxford University Press, 2012.

BRAZIL. Attorney General's Office (AGU). *Regulatory Sandbox Reference Guide*. Brasília, DF: Labori/AGU; Ministry of Development, Industry, Trade and Services (MDIC), 2024.

CAJUEIRO, Daniel Oliveira; CELESTINO, Victor Rafael Rezende. *A comprehensive review of Artificial Intelligence regulation*: Weighing ethical principles and innovation. Journal of Economy and Technology, 2025. doi:10.1016/j.ject.2025.07.001

COBRA, Thiago Guimarães de Barros. *Sandbox Regulatório: Regulação Experimental da Saúde.* Dissertation (Master's) — Faculdade de Direito, Universidade de São Paulo, São Paulo, 2024.

CORNELLI, Gabriele; FROST, Jon; GAMBACORTA, Leonardo; RAUCH, Michael; SONG, Zhongxia. Regulatory sandboxes and fintech funding: evidence from the UK. Basel: Bank for International Settlements, 2020. (BIS Working Papers, n. 901).

EBERS, Martin. Truly Risk-based Regulation of Artificial Intelligence: How to Implement the EU's AI Act. 2024.

EUROPEAN UNION. AI Office. *Third Draft of the General-Purpose AI Code of Practice*. 2025.

FINANCIAL CONDUCT AUTHORITY (FCA). Regulatory sandbox: lessons learned report. London: FCA, 2017.

GIKAY, Asress Adimi. Risks, innovation, and adaptability in the UK's incrementalism. 2024.

HANDRLICA, Jakub; SHARP, Vladimír; NEŠPOR, Jan. Forum shopping in regulatory sandboxes and the perils of experimental law-making. *Juridical Tribune – Tribuna Juridica*, v. 13, n. 3, p. 408–425, 2023. DOI: 10.24818/TBJ/2023/13/3.05. Available at: https://www.tribunajuridica.eu/arhiva/An13v3/5.%20Handrlica%2C%20Sharp%2C%20Nespor.pdf. Accessed on: 8 feb. 2026.

MAZO FORERO, Juan. Sandboxes regulatorios en inteligencia artificial. Bogotá: Universidad Externado de Colombia, Centro de Estudios sobre Genética y Derecho, 2025.

NIKOLINAKOS, Nikos. EU Policy and Legal Framework for AI. 2023.

NITI AAYOG. National Strategy for Artificial Intelligence (#AIForAll). New Delhi: Government of India, 2018.

ORGANISATION FOR ECONOMIC CO-OPERATION AND DEVELOPMENT (OECD). Recommendation of the Council on Artificial Intelligence (OECD/LEGAL/0449) – 2024 Update. 2024.

PORTELLA, Isabella Silva Di Jorge. Sandbox regulatório no Brasil: análise da implementação, dos modelos e da maturidade nas agências reguladoras federais. 2025. Professional Doctorate Thesis (Public Policy) – Escola Nacional de Administração Pública (Enap), Brasília, 2025.

ROBERTS, Huw; BABUTA, Alexander; MORLEY, Jessica; THOMAS, Christopher; TADDEO, Mariarosaria; FLORIDI, Luciano. *Artificial intelligence regulation in the United Kingdom*: a path to good governance and global leadership? Internet Policy Review, v. 12, n. 2, 2023. doi:10.14763/2023.2.1709

ROBLES-CARRILLO, Margarita. Sovereignty vs. Digital Sovereignty. Journal of Digital Technologies and Law, v. 1, n. 3, p. 673–690, 2023.

SCASSA, Teresa. Sovereignty and the Governance of Artificial Intelligence. UCLA Law Review Discourse, v. 71, p. 214–229, 2023–2024.

SOARES, Evandro da Silva. O sandbox como ambiente experimental de tecnologias emergentes, adaptação regulatória e inovação. Doctoral Thesis — Universidade Presbiteriana Mackenzie, São Paulo, 2025.

TRUBY, Jon; BROWN, Rafael; IBRAHIM, Imad Antoine; CAUDEVILLA, Oriol. *A sandbox approach to regulating high-risk artificial intelligence applications*. 12 nov. 2021. SSRN. Available at: https://ssrn.com/abstract=4373418. Accessed on: 12 fev. 2026.

UNITED NATIONS. General Assembly. Seizing the opportunities of safe, secure and trustworthy artificial intelligence systems for sustainable development. A/78/L.49, 21 Mar. 2024.

UNITED STATES. Executive Order 14110: Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence. Administration of Joseph R. Biden, Jr., 30 Oct. 2023.

UNITED STATES. International Trade Administration. South Korea Artificial Intelligence (AI) Basic Act. Washington, D.C.: U.S. Department of Commerce, 2025.

UNITED STATES. The White House Office of Science and Technology Policy (OSTP). Blueprint for an AI Bill of Rights: Making Automated Systems Work for the American People. Washington, DC, Oct. 2022.

UNITED STATES. The White House. Removing Barriers to American Leadership in Artificial Intelligence. Presidential Action, 23 Jan. 2025.

WORLD BANK. *Regulatory sandboxes for digital health*. Washington, DC: World Bank, 2025. Available at: https://documents1.worldbank.org/curated/en/099011825011040830/pdf/P175075-0c729174-1c7a-4f04-9bb2-4127f301c037.pdf. Accessed on: 12 fev. 2026.

ZHANG, Laney. China: Regulation of Artificial Intelligence. Washington, D.C.: Law Library of Congress, 2023.

III
COMPARATIVE AND GLOBAL PANORAMA: AMERICAS AND EUROPE

CANADA'S MIDDLE POWER AI REGULATION

Teresa Scassa
Canada Research Chair in Information Law and Policy
University of Ottawa

DOI: https://doi.org/10.59224/dti8.ch7

Abstract: This paper examines the evolving landscape of artificial intelligence (AI) regulation in Canada through the lens of middle-power governance in a rapidly shifting geopolitical environment. The paper argues that Canada's approach to AI regulation is shaped not only by external geopolitical pressures but also by internal constitutional and political realities. The article explores how concerns regarding sovereignty over digital infrastructure, data, and policy autonomy increasingly influence Canadian AI governance choices as it moves toward a hybrid model characterized by sector-specific regulation, soft law instruments, and international alliances.

Keywords: Artificial intelligence; sovereignty; middle power

Table of contents: 1. Context. 2. A Federal State with Obligations to Indigenous Communities. 2.1 Federalism. 2.2 Significant Indigenous Communities. 3. Sovereignty and AI. 3.1 Infrastructure. 3.2 Data. 3.3 Policy. 4. AI Regulation in Canada. 4.1 Federal law/policy. 4.1.1 Federal Public Sector. 4.1.2 Private Sector. 4.2 Provincial law/policy. 5. Conclusion.

On January 20, 2026, Canada's Prime Minister Mark Carney addressed the World Economic Forum in Davos and spoke of a "rupture in the world order". In his speech, he addressed the need of middle power countries to forge new alliances. He stated: "intermediate powers like Canada, are not powerless. They have the capacity to build

a new order that encompasses our values, like respect for human rights, sustainable development, solidarity, sovereignty, and territorial integrity of the various states."

Although the speech was about the changing global order and not artificial intelligence *per se*, artificial intelligence technologies will play a role in evolving global power dynamics. Many reports cite substantial potential economic gains from investment in and adoption of AI technologies,[1] although studies also warn that AI gaps will exacerbate existing inequalities among nations.[2] Embracing AI appears to be an economic must, yet regulating in a context where AI powers race to dominate is a challenge.

Since the start of the second Trump presidency in the United States, Canada's approach to AI regulation has noticeably shifted from one in which Canada framed itself as a leader in both AI development and regulation,[3] to one that is increasingly shaped by sovereignty concerns and trade considerations.[4] This paper traces the

---

1 Filippucci, F. et al. (2024), "The impact of Artificial Intelligence on productivity, distribution and growth: Key mechanisms, initial evidence and policy challenges", OECD Artificial Intelligence Papers, No. 15, OECD Publishing, Paris, https://doi.org/10.1787/8d900037-en; Annalisa Usardi and Bastien Drut, "The diverse economic impacts of artificial intelligence", SUERF Policy Brief No. 1174, 28 May 2025, https://www.suerf.org/publications/suerf-policy-notes-and-briefs/the-diverse-economic-impacts-of-artificial-intelligence/.

2 Eugenio M Cerutti, Antonio I Garcia Pascual, Yosuke Kido, Longji Li, Giovanni Melina, Marina Mendes Tavares, and Philippe Wingender. "The Global Impact of AI: Mind the Gap", IMF Working Papers 2025, 076 (2025), accessed 5/20/2026, https://doi.org/10.5089/9798229008570.001. See also: See, e.g., UN Trade and Development, "AI's $4.8 trillion future: UN Trade and Development alerts on divides, urges action", 7 April 2025, https://unctad.org/news/ais-48-trillion-future-un-trade-and-development-alerts-divides-urges-action.

3 See, e.g. Hon. François-Philippe Champagne, "Canada's Artificial Intelligence Revolution: Leading with Purpose", (2024), https://www.innovatingcanada.ca/technology/future-of-ai-2024/canadas-artificial-intelligence-revolution-leading-with-purpose/; Prime Minister of Canada, "Prime Minister highlights Canada's leadership in AI at the One Young World Summit in Montréal", 20 September 2024, https://www.pm.gc.ca/en/news/news-releases/2024/09/20/prime-minister-highlights-canadas-leadership-ai-one-young-world.

4 See, e.g.: Innovation, Science and Economic Development Canada, Canadian Sovereign AI Compute Strategy, 15 April 2026, https://ised-isde.canada.ca/site/ised/en/canadian-sovereign-ai-compute-strategy; Canada, Digital Sovereignty, 31 March 2026, https://www.canada.ca/en/shared-services/services/digital-sovereignty.html.

challenges for a country like Canada to regulate AI in the current global context and explores avenues that are open to it for regulating AI in accordance with its principles and values and perhaps through new alliances with middle power countries. Part I of this paper sets the context, explaining Canada's approach to AI regulation leading up to January 2025. Part II considers Canada's situation as a federal state as well as its obligations with respect to its Indigenous Peoples in shaping its approach to AI regulation and in building alliances. Part III examines digital sovereignty concerns and how these impact Canada's approach to AI regulation. Taking these diverse factors and considerations into account, Part IV provides an overview of the state of AI regulation in Canada and considers future directions.

1. CONTEXT

Canada has a track record of significant targeted investments in AI research and development, with the result that it has 'punched above its weight'[5] in the early days of the evolution of these technologies. Its Pan Canadian AI Strategy, launched in 2017,[6] was touted as the world's first national AI strategy.[7] It focused on the development of the AI sector in Canada, with investments in research, recruitment and

[5] There are frequent references to Canada 'punching above its weight' in AI research. See, e.g.: Universities Canada, "Canada's Universities: The source of Canada's global AI leadership", February 2026, https://univcan.ca/wp-content/uploads/2026/02/February-2026-The-source-of-Canadas-global-AI-leadership-Policy-primer.pdf, at p. 4; Dongwoo Kim, "Towards Canada's Asia Strategy on AI", Asia Pacific Foundation of Canada, October 8, 2019, https://www.asiapacific.ca/publication/towards-canadas-asia-strategy-ai;

[6] Canada, Industry, Science and Economic Development, "Pan-Canadian Artificial Intelligence Strategy", (2017), https://ised-isde.canada.ca/site/ised/en/pan-canadian-artificial-intelligence-strategy.

[7] CIFAR, "The Pan-Canadian AI Strategy," (n.d.), https://cifar.ca/ai/; Alberta Machine Intelligence Institute, "Pan-Canadian AI Strategy", (n.d.), https://www.amii.ca/about/pan-canadian-ai-strategy. See also: Hung, K.H., Li, L.X., Katrak, M. and Attard-Frost, B. (2025) "Canada's high-stakes artificial intelligence gamble: innovation policy, techno-nationalism and the political economy of jobs", Global Political Economy, 4(2): 256–269, DOI: 10.1332/26352257Y2025D000000040.

retention of talent.[8] In addition to its investments in the AI sector,[9] it has advanced several AI governance initiatives. For example, in 2019, Canada showed leadership in AI governance with its Directive on Automated Decision-Making for the federal public sector.[10]

In the early 2020s, Canada sought to leverage what it considered to be its global status in the field to push for harmonized and interoperable regulation of AI systems that adopted a principles-based approach.[11] In June 2022, when it introduced a proposed *Artificial Intelligence and Data Act* (AIDA) as Part III of Bill C-27,[12] it sought to align itself with the European Union on the one hand, with its draft *Artificial Intelligence Act*,[13] and a United States government on the other, which had issued

---

8 Hung, et al, *ibid*. at 258-259.

9 Government of Canada, "Canada moves toward safe and responsible artificial intelligence", 6 march 2025, https://www.canada.ca/en/innovation-science-economic-development/news/2025/03/canada-moves-toward-safe-and-responsible-artificial-intelligence.html.

10 Canada, Directive on Automated Decision Making (2019), https://www.tbs-sct.canada.ca/pol/doc-eng.aspx?id=32592&section=html#appA. See also:

11 Key principles were outlined in Canada's Digital Charter, which was announced in 2019: Canada, Canada's Digital Charter (2019), https://ised-isde.canada.ca/site/innovation-better-canada/en/canadas-digital-charter-trust-digital-world; Teresa Scassa, "Administrative Law and the Governance of Automated Decision-Making: A Critical Look at Canada's Directive on Automated Decision-Making", (2021) 54:1 University of British Columbia Law Review 251-298, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3722192.

12 Bill C-27, An Act to enact the Consumer Privacy Protection Act, the Personal Information and Data Protection Tribunal Act and the Artificial Intelligence and Data Act and to make consequential and related amendments to other Acts, 1st Sess, 44th Parl, 2022 (first reading 16 June 2022), online: <www.parl.ca/DocumentViewer/en/44-1/bill/C-27/first-reading> [perma.cc/] [Bill C-27]. Note that the short title of Bill C-27 is Digital Charter Implementation Act, 2022. Bill C-27 consists of the Consumer Protection Privacy Act [CPPA], Personal Information and Data Protection Tribunal Act, and Artificial Intelligence and Data Act [AIDA].

13 "Proposal for a Regulation of The European Parliament and of the Council Laying Down Harmonised Rules On Artificial Intelligence (Artificial Intelligence Act) And Amending Certain Union Legislative Acts" (2021), online: *EUR-Lex* <eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52021PC0206> [perma.cc/]. This later became the EU Artificial Intelligence Act, Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and

an Executive Order supporting principles-based AI development in the US public sector,[14] combined with support for the NIST *Risk Management Framework*,[15] which proposed a voluntary risk-based approach to AI governance for organizations. In February 2025, Canada signed the Council of Europe's *Framework Convention on Artificial Intelligence*,[16] embracing a human-rights based approach to AI governance, particularly for the public sector.

The second Trump presidency, which began in January 2025 brought significant changes to Canada's approach to AI regulation. These changes are rooted in a starkly different relationship between Canada its neighbor to the south. Persistent comments from the US president about Canada as the 51st US state, combined with a clear Trump expansionist agenda, have raised significant sovereignty concerns. Punishing tariffs and economic threats have also impacted Canadian policy approaches, causing the Canadian government to dramatically shift its economic and policy priorities. Both digital and data sovereignty have become important preoccupations for Canada, as the country confronts the reality that much of its technological infrastructure is subject to US control.

The dramatic turn by the Trump administration away from principles-based AI governance and towards a race for AI domination[17] has also led Canada to retreat from more prescriptive European-style risk-regulation-based governance of AI. This

Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act).

14 "Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence", US Executive Order 14110 of October 30, 2023, https://www.federalregister.gov/documents/2023/11/01/2023-24283/safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence.

15 National Institute of Standards and Technology, "Artificial Intelligence Risk Management Framework (AI RMF 1.0)" (January 2023), online (pdf): <www.nist.gov/itl/ai-risk-management-framework> [perma.cc/] [NIST Framework].

16 2024 Council of Europe Framework Convention on Artificial Intelligence, 133rd Session of the Committee of Ministers (Strasbourg, 17 May 2024), https://www.coe.int/en/web/artificial-intelligence/the-framework-convention-on-artificial-intelligence.

17 See: The White House, Executive Order 14179 – Removing Barriers to American Leadership in Artificial Intelligence, January 23, 2025, https://www.whitehouse.gov/presidential-actions/2025/01/removing-barriers-to-american-leadership-in-artificial-intelligence/. This order states: "It is the policy of the United States to sustain and enhance America's global AI dominance in order to promote human flourishing, economic competitiveness, and national security."

may be at least in part out of concerns over potential tariff-based retaliation.[18] Canada's proximity to the US also creates a risk of loss of talent and innovation from Canada to the US.[19] Some attempts by Canada to address digital issues linked to US platforms have led to resistance from the US, and threats of tariffs and other retaliation contribute to concerns over the potential economic consequences of 'excessive' regulation.[20] Canada's Minister of Artificial Intelligence and Digital Innovation has referred to the government's new goal of promoting AI regulation that is "light, tight and right".[21]

In spite of the upheaval in Canada's approach to AI regulation, some form of regulation is required. Although Canada has ambitious plans for adoption of AI to advance its own innovation and productivity agendas, trust in AI remains significantly lower in Canada than in comparable nations.[22] Concerns over the potential harms of AI technologies continue to rise, particularly as high profile issues grip media attention.[23] Impacts on the labour market and the environment raise concerns

---

18 See, e.g., Michael Geist, "Why the Trump Trade Threats Will Place Canadian Digital, Cultural, and AI Policy Under Pressure", February 23, 2025, https://mgeist.substack.com/p/why-the-trump-trade-threats-will.

19 See, e.g., Lawrence Zhang and Daniel Castro, "Canada should harness its AI advantage, not squander it", Macdonald Laurier Institute, May 19, 2026, https://macdonaldlaurier.ca/canada-should-harness-its-ai-advantage-not-squander-it-lawrence-zhang-and-daniel-castro/; Rami Alhamad, "Canada Is Spending Billions on AI. Why Are Companies Still Fleeing?", The Walrus, 20 May 2026, https://thewalrus.ca/canada-is-spending-billions-on-ai-why-are-companies-still-fleeing/.

20 See Geist, *supra* note 18. See discussion Infra in Part III(C).

21 Kate Lunau, "Evan Solomon Wants Canada to Trust AI. Can We Trust Evan Solomon?", The Walrus, January 15, 2026, https://thewalrus.ca/evan-solomon-ai/.

22 See, e.g., N. Gillespie, S. Lockey, T. Ward, A. Macdade, & G. Hassed, (2025). Trust, attitudes and use of artificial intelligence: A global study 2025. The University of Melbourne and KPMG. DOI 10.26188/28822919; Eddie Sheppard and David Coletto, "What do Canadians think about AI: A Tool for Progress or a Threat to Stability?", Abacus Data, 19 September 2025, https://abacusdata.ca/what-do-canadians-think-about-ai-a-tool-for-progress-or-a-threat-to-stability/.

23 Some of these issues include the harms resulting from youth reliance on AI companions (see, e.g.: Rhitu Chatterjee, "their teenage sons died by suicide. Now, they are sounding an alarm about AI chatbots", NPR online, 19 September 2025, https://www.npr.org/sections/shots-health-news/2025/09/19/nx-s1-5545749/ai-chatbots-safety-openai-meta-characterai-teens-suicide), as

that go beyond AI safety.[24] Canada has identified six pillars for its next AI Strategy.[25] In addition to highlighting the need to build new trusted partnerships and global alliances, one of the pillars relates directly to the meat of AI regulation: "protecting Canadians and safeguarding our democracy".[26]

2. A FEDERAL STATE WITH OBLIGATIONS TO INDIGENOUS COMMUNITIES

Apart from the broader geopolitical context in which it finds itself, Canada's approach to AI regulation is shaped by its status as a federal state, which is both linked to its colonial past and embedded in its current political realities. It is also on a pathway to reconciliation with its Indigenous Peoples. What unites these two themes is the need for dialogue and engagement to chart a constructive path forward

well as the controversy over the link between use of OpenAI's Chatbot and a mass shooting tragedy in Canada (see, e.g.: Nadine Yousif, "Seven lawsuits filed against OpenAI by families of Canada mass-shooting victims", BBC online, 29 April 2026, https://www.bbc.com/news/articles/c99l03k0ly4o.)

[24] See, e.g. Adam Zewe, "Explained: Generative AI's environmental impact", MIT News, 17 January 2025, https://news.mit.edu/2025/explained-generative-ai-environmental-impact-0117; Apoorva Chouksey, Aneha K. Rajan, Vikas Gurjar, Rajnarayan Tiwari, Pradyumna Kumar Mishra, "The green paradox: The climate, environmental, and sustainability implications of artificial intelligence", (2026) 6 Global Environmental Change Advances 100029, https://doi.org/10.1016/j.gecadv.2025.100029; Economic Development Research Partners, "Artificial Intelligence Impact on Labor Markets: Literature Review", International Economic Development Council, 13 March 2025, https://www.iedconline.org/clientuploads/EDRP%20Logos/AI_Impact_on_Labor_Markets.pdf; Tahsin Mehdi and Marc Frenette, "Exposure to artificial intelligence in Canadian jobs: Experimental estimates", Statistics Canada, 25 September 2024, https://www150.statcan.gc.ca/n1/pub/36-28-0001/2024009/article/00004-eng.htm.

[25] These include: "Protecting Canadians and safeguarding our democracy; Empowering Canadians; Powering AI adoption for shared prosperity; Building the Canadian sovereign AI foundation; Scaling Canadian champions; Building trusted partnerships and global alliances." See: Benjamin Lopez Steven, "Feds reveal 6 pillars for long-touted, repeatedly delayed national AI strategy", CBC News, April 28, 2026, https://www.cbc.ca/news/politics/ai-strategy-pillars-evan-solomon-9.7180418.

[26] *Ibid*.

and a need to respect diversity.

Benz and Sonnicksen observe that the structure of a federal state can enable greater "'choice' and 'voice'", but can also "cause deadlocks or obstruct governability in a democracy".[27] They describe the division of powers in federal states as creating a conflicting relationship "which at its roots, relates to the problem of accommodating collective and individual self-determination and the continuous balancing act between unity and diversity."[28] These features of collaboration and dialogue may ultimately be important skills in building and maintaining middle power alliances and for charting a new path for principled AI governance.

2.1 Federalism

Canada is a federal state consisting of ten provinces and three territories. Under the *Constitution Act, 1867*,[29] legislative powers are divided between the federal and provincial governments. The Supreme Court of Canada has referred to federalism as "the political mechanism by which diversity could be reconciled with unity",[30] and the choice of a federal structure for Canada is tied to its colonial history and its diverse populations across a vast and varied geography. Different trends in constitutional interpretation over the country's history have ranged from a tendency by the Judicial Committee of the Privy Council to give greater power to the regions,[31] to a shift towards more centralized power.[32] More recently, the Supreme Court of Canada has favoured what it calls co-operative federalism – a doctrine that recognizes that some issues have both federal and provincial dimensions and that jurisdiction over

---

27 Arthur Benz and Jared Sonnicksen, "Patterns of federal democracy: tensions, friction, or balance between two government dimensions", European Political Science Review (2017), 9:1, 3–25, at 6.

28 *Ibid*.

29 *The Constitution Act, 1867*, 30 & 31 Vict, c 3.

30 Reference re Employment Insurance Act (Can.), ss. 22 and 23, [2005] S.C.J. No. 57, 2005 SCC 56, [2005] 2 S.C.R. 669, at para. 10 (S.C.C.)

31 See, e.g.: A.C. Cairns, "The Judicial Committee and Its Critics" (1971) 4 Canadian Journal of Political Science 301

32 See: Katherine Swinton, "Federalism Under Fire: The Role of the Supreme Court of Canada" (1992) 55 Law and Contemporary Propblems 121, at 124.

certain subject matters can operate in tandem.[33] Newman describes this as "perhaps less a principle than a practice, and more political than legal in its nature and substance".[34] It encourages federal and provincial governments to work together towards common goals in areas where legislative competence can overlap. As the Supreme Court of Canada has stated, "Co-operative federalism reflects the realities of an increasingly complex society that requires the enactment of co-ordinated federal and provincial legislative schemes to better deal with the local needs of unity and diversity".[35]

Emerging technologies have posed challenges to the constitutional division of powers in part because they were unanticipated at the time of confederation and in part because they can have broad effects that cut across jurisdictional lines. Data protection offers a good example. At a time when personal information was recorded and stored on paper or on largely unconnected digital systems, data was considered a by-product of public and private sector activities that were governed by the federal or provincial jurisdiction into which the subject matter that generated the data fell. As a result, Canada has fourteen public sector data protection laws (one federal, ten provincial and three territorial), four private sector data protection laws (one federal and one in each of Alberta, British Columbia and Quebec), as well as health-sector specific data protection laws in most provinces and territories.[36] While the laws share similarities, they are not harmonized, creating a patchwork data protection framework that has proven challenging for data interoperability and mobility in a digital context.[37] With access to data for AI innovation being an important current issue, Canada's fragmented data protection landscape is less than ideal.

In some cases, different aspects of complex issues may fall within federal jurisdiction and others within provincial. For example, in Canada, competition law is a

---

33 See: Warren J. Newman, "The Promise and Limits of Cooperative Federalism as a Constitutional Principle", (2016) 76 Supreme Court Law Review (2d) 67.

34 Ibid. at 71.

35 Quebec (Attorney General) v. Canada (Attorney General), [2015] S.C.J. No. 14, 2015 SCC 14, [2015] 1 S.C.R. 693 (S.C.C.), at para 148.

36 Barbara von Tigerstrom, *Information and Privacy Law in Canada*, (Irwin Law, 2020), at 444-45.

37 See, e.g., Competition Bureau, Unlocking the power of health data, June 23, 2022, https://competition-bureau.canada.ca/en/unlocking-power-health-data.

matter of federal jurisdiction, while the provinces govern consumer protection law. The issue of algorithmic pricing, for example, may have both competition (federal jurisdiction) and consumer protection (provincial jurisdiction) dimensions.[38]

Another example where the federal division of power renders AI regulation complex is in the health sector. The publicly funded health care system is administered at a provincial level, and privacy/data protection for health data collected through that system is also a matter of provincial jurisdiction. Health care professionals and hospitals are provincially regulated. At the same time, the regulation of medical devices (including those involving AI technologies) falls under federal jurisdiction, largely supported by the federal criminal law power.[39] The federal government can also influence some aspects of health data policy through its overall spending power.[40] Both the provinces and the federal government have public health mandates.[41] From the perspective of AI regulation in healthcare, these different areas of legislative competence create a complex landscape.

Among the world's twenty-five federal states are several middle power states, including Brazil, Germany, Mexico, India and Argentina. Each federal state is different in terms of the way in which legislative power is shared between national and subnational governments. Federal states may find it more complex to regulate AI and to do so in a comprehensive and systematic fashion. The degree of difficulty may be

---

38 The federal Competition Bureau released a discussion paper on Algorithmic Pricing. See: Competition Bureau of Canada, "Algorithmic pricing and competition: Discussion paper", 10 June 2025, https://competition-bureau.canada.ca/en/how-we-foster-competition/education-and-outreach/publications/algorithmic-pricing-and-competition-discussion-paper. In the meantime, the Government of Manitoba has introduced a bill to prohibit algorithmic pricing under its business practices legislation: Bill 49, The Business Practices Amendment Act, 3rd Sess., 43rd Legislature, 2026, https://web2.gov.mb.ca/bills/43-3/b049e.php.

39 Martha Butler and Marlisa Tiedemann, The Federal Role in Health and Health Care, Library of Parliament, 2011, Publication No. 2011-91-E, https://lop.parl.ca/sites/PublicWebsite/default/en_CA/ResearchPublications/201191E.

40 See: Howard Leeson, "Constitutional Jurisdiction Over Health and Health Care Services in Canada", Discussion Paper No. 12, Government of Canada, August 2022, https://publications.gc.ca/collections/Collection/CP32-79-12-2002E.pdf/

41 Keri Gammon, "Pandemics and Pandemonium: Constitutional Jurisdiction Over Public Health", 2006 15 Dalhousie Journal of Legal Studies 1.

impacted by the nature of the constitutional division of powers, with some federal states having far more centralized powers than others. Where regions are more autonomous, there may be less capacity for broad-based national action. This can generate complexities in an environment where international harmonization of rules and approaches is considered desirable to facilitate trade. In an area that is as fast moving as AI, and that has such significant international dimensions, the inability of a central authority to regulate comprehensively and to do so swiftly can be a challenge. It can also mean that within a single nation, different regions have different levels of protection regarding the use of AI systems.

Nevertheless, federalism is also a mode for managing diversity within a broader political and economic framework. This capacity for diversity and its management could have unexpected benefits for federal states when it comes to technology regulation. While the ability to respond rapidly to a powerful new technology may be seen as important, speed of action can sometimes lead to the adoption of premature or flawed solutions. Further, differences in sub-national approaches can allow for the exploration of diverse models and could, in theory, lead to eventual harmonization around more successful approaches. Nevertheless, without strong incentives and good frameworks for collaboration and harmonization, fragmentation and internal trade barriers can be the result.

In practical terms, AI regulation in a federal state is likely to be slower and more decentralized. Federal governments can play a convening or coordinating role, although domestic politics may impact the extent to which this is possible. Because federalism may inhibit comprehensive centralized regulation, federal states may also come to rely more heavily on a combination of corporate self-governance and ex-post litigation. Sectoral regulation may be either favoured over or more likely to succeed than comprehensive risk-regulation legislation. When it comes to the type of cross-sectoral AI risk regulation found in the EU *AI Act*, it is clear that such an approach is significantly constrained in a federation such as Canada's. The challenge of regulating AI in a federal state without strong centralized powers is to adapt to what is possible through a combination of the exercise of legislative competence, persuasion, and collaboration.

2.2 Significant Indigenous Communities

As a nation founded through colonization, Canada has significant Indigenous communities. These communities fall into three broad categories: First Nations, Inuit and Métis (FNIM). Although there is collaboration and cooperation among FNIM Peoples at the collective level, local communities exist with varying degrees of autonomy under different treaties or other arrangements on pathways to self-determination. FNIM Peoples have struggled under often brutal colonial policies, attempts at forced assimilation, racism, neglect and impoverishment. A Truth and Reconciliation process that culminated in 2015 with a series of calls to action has slowly gathered some momentum at different levels of government.[42] Ongoing land claims processes have also contributed to the growing autonomy of some communities. Canada's ratification of the *United Nations Declaration on the Rights of Indigenous Peoples* (UNDRIP)[43] has also strengthened commitments to work in partnership with FNIM Peoples, particularly with the implementation of this treaty in 2021.[44]

Indigenous Data Sovereignty (IDS) has become an important part of the movement towards self-determination of Canada's FNIM Peoples. IDS is defined as "the right of Indigenous peoples to govern the collection, ownership, and application of data about Indigenous communities, peoples, lands, and resources",[45] and efforts in

---

42 Canada, Truth and Reconciliation Commission of Canada. Canada's residential schools: the final report of the Truth and Reconciliation Commission of Canada (2015), https://publications.gc.ca/site/eng/9.807830/publication.html. See also: National Centre for Truth and Reconciliation, Truth and Reconciliation Commission of Canada: Calls to Action, https://nctr.ca/about/truth-and-reconciliation-commission-of-canada-calls-to-action/.

43 UN General Assembly, United Nations Declaration on the Rights of Indigenous Peoples : resolution / adopted by the General Assembly, A/RES/61/295, 2 October 2007, https://www.refworld.org/legal/resolution/unga/2007/49353.

44 *United Nations Declaration on the Rights of Indigenous Peoples Act*, SC 2021, c. 14.

45 Stephanie Carroll Rainie, Tahu Kukutai, Maggie Walter, Oscar Luis Figueroa-Rodriguez, Jennifer Walker & Per Axelsson, Indigenous Data Sovereignty in THE STATE OF OPEN DATA 300, 300–19 (Davies et al. ed., 2019) at 301.

Canada are part of a broader international movement.[46] The First Nations Information Governance Council (FNIGC) has led the way in Canada with its development of the OCAP principles that have served to define an approach to data sovereignty based on principles of ownership, control, access and possession.[47] Métis peoples have also developed their own data strategy, which is founded upon sovereignty concepts,[48] and Inuit are in the process of doing so, following the articulation of a research policy.[49] For many FNIM communities, sovereignty over health data is a priority, given the significant health challenges faced by these communities. Because health data is primarily a matter of provincial concern, FNIM peoples been working with provincial governments to develop sovereign approaches to Indigenous health data governance,[50] although there remains much to be done and work is ongoing.

Although provinces clearly have a role to play in supporting IDS, the federal government is also a key player as it has both important constitutional jurisdiction over FNIM peoples but also represents the Crown in the context of treaty and Indigenous rights. At the federal level, a recent discussion paper on modernizing the federal public sector data protection law – the *Privacy Act*[51] – features a set of proposed reforms that would recognize and take certain steps towards IDS at the federal level.[52] The document speaks of "partnership with Indigenous Peoples to ensure that

---

46 See, e.g.: Maggie Walter, Tahu Kukutai, Stephanie Russo Carroll and Desi Rodriguez-Lonebear, eds., Indigenous Data Sovereignty and Policy, Taylor Francis, 2021.

47 First Nations Information Governance Centre, Ownership, Control, Access and Possession (OCAP™) (2014).

48 Métis National Council, *Métis National Data Strategy*, 15 October 2015, https://www.metisnation.ca/wp-content/uploads/2025/12/Consolidated_Me_tis_Nation_Data_Strategy_Final_15_10_2025_3_56.pdf, at 11-12.

49 Inuit Tapiriit Kanatami, National Inuit Strategy on Research, 2018, https://itk.ca/projects/national-inuit-strategy-on-research/.

50 See: Indigenous Population Health: Our Data, Our Stories, Indigenous Primary Health Care Council, 2025, https://iphcc.ca/wp-content/uploads/ninja-forms/2/D4-Data-Governance-FINAL-1.pdf.

51 RSC 1985, c P-21.

52 Treasury Board Secretariat, 2026 Review of the Privacy Act: Policy Approaches, 2 April 2026, https://www.canada.ca/en/treasury-board-secretariat/services/access-information-privacy/access-information/privacy-act-modernization-policy-approaches/2026-review-privacy-act-policy-approaches.html [TBS Review]. The recommendations flow from two rounds of engagement

privacy rules support their governance needs."[53] The discussion paper proposes creating a new category of "Indigenous personal data", which would be identified as such with the consent of the data subject, and which would be shared with the relevant Indigenous governments under a data sharing agreement.[54]

Although considerable work has been done to enable IDS, there has been less attention paid to AI governance and Indigenous communities. Nonetheless, AI issues will be important, impacting control and consent, how fragile languages are included in training data, the representation of Indigenous peoples and world views in training data, and the ingestion of culturally sensitive content into generative AI systems.[55] Other issues will include addressing bias and discrimination in AI systems. The issues may be different depending on whether AI systems are controlled or closed systems and systems, and whether they are built with trusted partners. There will be significant issues for art and culture, traditional knowledge, and language. Work is already underway to explore the potential for Indigenous-led development of both large and small language models that respect IDS principles in their incorporation of content.[56]

Canada is not the only nation with significant Indigenous communities, and indeed, some of the AI-related issues facing Indigenous communities are aligned

with Indigenous partners: Report on Engagement with Indigenous Partners and Report on 2022 Engagement with Indigenous Partners https://www.justice.gc.ca/eng/csj-sjc/pa-lprp/wwh4-cqnae4/index.html.

53 TBS Review, *supra* note 52 at Theme 5.

54 *Ibid*., Proposal 17.

55 Gin Sexsmith, "AI can help protected threatened languages – but it can also exploit them", Canadian Geographic, 20 April 2026, https://canadiangeographic.ca/articles/ai-can-help-protect-threatened-languages-but-it-can-also-exploit-them/. See also the work being done by the Abundant Intelligences project of Indigenous AI, https://www.indigenous-ai.net/abundant/.

56 Ashley Goodfellow Craig, "How York U is using technology to protect Indigenous languages", YorkU, 6 November 2025, https://www.yorku.ca/yfile/2025/11/06/how-york-u-is-using-technology-to-protect-indigenous-languages/; Brooke Tanner and Cameron F. Kerry, "Can small language models revitalize Indigenous languages?", Brookings, 19 March 2025, https://www.brookings.edu/articles/can-small-language-models-revitalize-indigenous-languages/; Anna Dorn, "How AI can adapt to Indigenous knowledge", Medium, 7 November 2025, https://medium.com/blog/how-ai-can-adapt-to-indigenous-knowledge-2ae3cd442096.

with concerns facing other middle power countries with important Indigenous communities or with cultural and linguistic minorities facing similar issues of inclusion and empowerment with respect to AI technologies.

3. SOVEREIGNTY AND AI

Canada, like many other middle-power nations, is confronting significant digital sovereignty issues when it comes to AI. The term 'digital sovereignty' is not new, although it is capable of multiple meanings. Kristina Irion describes digital sovereignty as "conceptually fuzzy", and she notes that it is "used to animate a wide spectrum of geopolitical, normative and industrial ambitions."[57] In this paper, "digital sovereignty" is used as a broad category, subdivided into issues of sovereignty over infrastructure, data, and digital law and policy.

The issues considered in this section are framed as national sovereignty issues, but in the context of federalism, it is important to remember that sub-national governments may be sovereign with respect to their areas of exclusive legislative competence. Further, the claims to data sovereignty asserted by Indigenous peoples, discussed above, are dimensions of broader claims to self-determination. It is important, therefore, to keep in mind that there may be multiple layers of sovereignty issues.[58]

3.1 Infrastructure

Since the change in relationship with the United States, concerns over the sovereignty of digital infrastructure in Canada abound. Sovereignty issues linked to infrastructure include things such as being able to own and control internet and telecommunications infrastructure, cloud-based services, and even the systems used on a

---

[57] Kristina Irion, "The General Data Protection Regulation though the lens of digital sovereignty", EU Cyber Direct Research Seminar: EU Digital Sovereignty: From Narrative to Policy, 18 March 2022, https://www.ivir.nl/publicaties/download/Irion-GDPR-and-Digital-Sovereignty-11Mar22.pdf, at 1.

[58] See, e.g., Teresa Scassa, "Sovereignty and the Governance of Artificial Intelligence", (2024) 71 UCLA L. Rev. Disc. 214-229.

daily basis for government, business and educational services.[59] The routing of internet communications in Canada relies on US infrastructure,[60] and Canadians are heavily dependent on US-based social media platforms for news and other forms of communication.[61] It can also include, in the AI context, the sovereign computing power necessary to independently develop large and complex AI systems.

The potential impacts of a lack of sovereignty over digital infrastructure are diverse and include concerns over the potential of infrastructure to be used for espionage or to secure access to data by foreign governments, [62] for it to be used to manipulate populations,[63] or for it simply to be turned off or disrupted.[64] Underlying the infrastructural concerns are other sovereignty-related preoccupations, including the ability to foster or protect national production and distribution of culture,[65] to develop AI tools responsive to national or local conditions.[66] Concerns have also been

---

59 See, e.g. Heidi Tworek and Alicia Wanless, "How to Confront Canada's Digital Dependence", 1 July 2025, Centre for International Governance Innovation, https://www.cigionline.org/articles/how-to-confront-canadas-digital-dependence/.

60 See, e.g., Andrew Clement, "Canadian Network Sovereignty: A Strategy for Twenty-First-Century National Infrastructure Building", 26 March 2018, CIGI Online, https://www.cigionline.org/articles/canadian-network-sovereignty/.

61 For concerns over Canada's dependence on these US-based platforms, see: Michael Karanicolas, "U.S. social media platforms are a national security threat", Policy Options, 8 March 2025, https://policyoptions.irpp.org/2025/03/social-media-threat/; Alexander Martin, "The threat of US-centric social media", Open Canada, 31 March 2026, https://opencanada.org/the-threat-of-us-centric-social-media/.

62 See, e.g.: Kadri Kaska, Henrik Beckvard and Tomás Minárik, "Huawei, 5G and China as a Security Threat", CCDCOE, 2019, https://ccdcoe.org/uploads/2019/03/CCDCOE-Huawei-2019-03-28-FINAL.pdf.

63 Andrew Iliadis, Francesca Tripodi, Aashka Dave, Leslie Kay Jones, Amelia Acker and Heather Ford, "Infrastructures of Manipulation", 2023 AoIR2023, https://spir.aoir.org/ojs/index.php/spir/article/view/13539.

64 See *supra*, note 70.

65 See, e.g., UNESCO, Report of the Independent Expert Group on Artificial Intelligence and Culture, 2025, https://www.unesco.org/sites/default/files/medias/fichiers/2025/09/CULTAI_Report%20of%20the%20Independent%20Expert%20Group%20on%20Artificial%20Intelligence%20and%20Culture%20%28final%20online%20version%29%201.pdf.

66 See, e.g., Alexander Martin, "The AI Threat to Canadian Democracy: Fighting for Digital Sovereignty", Open Canada, 26 May 2025, https://opencanada.org/the-ai-threat-to-canadian-democ-

raised regarding the 'software stack' that forms part of softer digital infrastructure.[67]

As part of its overall AI strategy, Canada has committed to investing in sovereign AI computing capability.[68] It is also developing middle-power international partnerships in this regard. A recent joint declaration with Germany emphasised "both countries' shared commitment to advancing secure, resilient and sovereign AI capabilities in a rapidly evolving global landscape".[69] This is alongside other international

racy-fighting-for-digital-sovereignty/, Angelo Toma, "Algorithmic Foreign Influence: Rethinking Sovereignty in the Age of AI", Lawfare, 20 August 2025, https://www.lawfaremedia.org/article/algorithmic-foreign-influence--rethinking-sovereignty-in-the-age-of-ai.

67 It should be noted that the "software stack" that runs on digital infrastructure in Canada is also heavily US owned. (Catherine McIntyre, "New data shows just how dependent Canadian businesses are on U.S. tech", The Logic, 10 March 3036, https://thelogic.co/news/canada-tech-sovereignty-buy-canadian/.) Concerns raised in this context include the ability of US-based companies to "turn off" access to these technologies if ordered to do so. Precedent for this was seen in February 2025, when the Trump administration imposed sanctions against the International Criminal Court and some of its members. (White House, Executive Order 14203, Imposing Sanctions on the International Criminal Court, 6 February 2025, https://www.whitehouse.gov/presidential-actions/2025/02/imposing-sanctions-on-the-international-criminal-court/.) The sanctions were in relation to preliminary investigations "concerning personnel of the United States and certain of its allies, including Israel". Among other things, the sanctions led to the affected individuals being cut off from use of their credit cards, Amazon accounts and other services. (Naomi O'Leary, "'It's surreal': US sanctions lock International Criminal Court judge out of daily life", The Irish Times, 12 December 2025, https://www.irishtimes.com/world/us/2025/12/12/its-surreal-us-sanctions-lock-international-criminal-court-judge-out-of-daily-life/.) Note that in the EU, governments are planning moves away from US-based Big Tech and towards open source solutions. See: Matt Burgess and Vittoria Elliot, "The EU is Going Through a Trump-Fueled Breakup with Big Tech", WIRED, 21 May 2026, https://www.wired.com/story/the-eu-is-going-through-a-trump-fueled-breakup-with-big-tech/.

68 Canada, "Canada launches national initiative to build large-scale AI supercomputing capacity", 15 April 2026, https://www.canada.ca/en/innovation-science-economic-development/news/2026/04/canada-launches-national-initiative-to-build-large-scale-ai-supercomputing-capacity.html; Innovation, Science, and Economic Development, "Canadian Sovereign AI Compute Strategy", 15 April 2026, https://ised-isde.canada.ca/site/ised/en/canadian-sovereign-ai-compute-strategy.

69 Innovation, Science and Economic Development Canada, "Canada and Germany sign AI joint declaration and launch Sovereign Technology Alliance", 14 February 2026, https://www.canada.ca/en/innovation-science-economic-development/news/2026/02/canada-and-germany-sign-

agreements with the UK, Spain, and the EU which touch on digital sovereignty issues and AI.[70] Canada is also investing in the building of data centres as part of a commitment to developing sovereign infrastructure.[71]

Canada is not the only country to turn its attention to digital sovereignty issues. Nevertheless, sovereignty in this context can pose some interesting challenges, as localization of digital and data services can act as important trade barriers, particularly in an increasingly digital world. The US has signaled that moves by other countries to assert greater digital sovereignty will be regarded as trade irritants.[72]

3.2 Data

Data sovereignty issues are not new, although they have become more acute over time. Hummel et al suggest that "Data sovereignty typically relates in some way to

ai-joint-declaration-and-launch-sovereign-technology-alliance.html.

70 Innovation, Science and Economic Development Canada, "Canada advances international cooperation on industry, AI and digital technology at the G7 Industry, Digital and Technology Ministers' Meeting", 9 December 2025, https://www.canada.ca/en/innovation-science-economic-development/news/2025/12/canada-advances-international-cooperation-on-industry-ai-and-digital-technology-at-the-g7-industry-digital-and-technology-ministers-meeting.html; Innovation, Science and Economic Development Canada, "Minister Solomon, His Majesty King Felipe VI of Spain and Spanish Deputy Prime Minister Cuerpo advance Canada-Spain artificial intelligence cooperation through new memorandum of understanding", 20 May 2026, https://www.newswire.ca/news-releases/minister-solomon-his-majesty-king-felipe-vi-of-spain-and-spanish-deputy-prime-minister-cuerpo-advance-canada-spain-artificial-intelligence-cooperation-through-new-memorandum-of-understanding-856857718.html.

71 Innovation, Science and Economic Development, Canada, "Enabling large-scale sovereign AI data centres", 16 February 2026, https://ised-isde.canada.ca/site/ised/en/enabling-large-scale-sovereign-ai-data-centres.

72 See, e.g., Michael Geist, "The US isn't happy about Canada's quest for digital sovereignty", 9 April 2026, Globe and Mail, https://www.theglobeandmail.com/business/commentary/article-the-us-isnt-happy-about-canadas-quest-for-digital-sovereignty/; Raphael Satter and Alexandra Alper, "Exclusive: US orders diplomats to fight data sovereignty initiatives", Reuters, 25 February 2026, https://www.reuters.com/sustainability/boards-policy-regulation/us-orders-diplomats-fight-data-sovereignty-initiatives-2026-02-25/.

meaningful control, ownership, and other claims to data or data infrastructures,"[73] although they note that different users give the concept different meaning.

In 2012, Irion wrote about the sovereignty challenges presented by the move by governments to cloud services. Acknowledging the potential for the term to take on multiple meanings, she defined data sovereignty as: "[g]overnment's exclusive authority and control over all virtual public assets, which are not in the public domain, irrespective whether they are stored on their own or third parties' facilities and premises."[74] These virtual public assets include personal data held by governments about their citizens and residents, but also include confidential commercial information and trade secrets shared with government, as well as data important to national security. Control over these core data is essential to data sovereignty. As Irion notes, "If crucial information were controlled by a foreign power the country would be a colony."[75]

In Canada, a 2018 a federal government policy paper on moving to largely US-owned cloud services contained little discussion of sovereignty concerns, perhaps reflecting a rather complacent approach to its relationship with that country at the time.[76] That changed in January of 2025, with the actions of the new US regime leading to considerable concerns over the vulnerability of Canada due to its dependence on US technological infrastructure including cloud data storage.[77]

---

[73] Patrick Hummel, Matthias Braun, max Tretter and Peter Dabrock, "Data sovereignty: A review", (2021) Big Data and Society 1, https://d-nb.info/1282975056/34, at 2.

[74] Kristina Irion, "Government Cloud Computing and National Data Sovereignty", (2021) 4:3-4 Policy & Internet 40, at 41.

[75] *Ibid*. at 53.

[76] Government of Canada Cloud Adoption Strategy, 2018, https://www.canada.ca/en/government/system/digital-government/digital-government-innovations/cloud-services/government-canada-cloud-adoption-strategy.html.

The document did mention security and localization issues but in doing so referred to existing policy documents on cloud services and data residency.

[77] An updated document addressed cloud security risk management in much greater detail. See: Government of Canada Cloud Security Risk Management Approach and Procedures, 22 October 2025, https://www.canada.ca/en/government/system/digital-government/digital-government-innovations/cloud-services/cloud-security-risk-management-approach-procedures.html. For a critique of the government's approach, see: Lawrence Zhang, "The right way for Canada to secure

The solution to data sovereignty concerns has not necessarily been to move away from cloud storage, particularly as the cloud can offer important efficiency and security advantages. Although contractual arrangements between data holders and cloud service providers typically establish a high level of confidentiality and data security, these are private law arrangements, and the risk remains that foreign states can still compel cloud providers to share data in specific circumstances. In Canada, a major concern has been the US *CLOUD Act*[78] and the powers it provides US authorities to access data held by US companies. Complicating the situation is the application of the law to companies with a range of different connections to the US, which can include major Canadian companies.[79] Such concerns have spawned more recent calls for data localization of more sensitive Canadian data, including personal health information, as well as calls for encryption of sensitive cloud-stored data.[80]

Many equate data localization (the retention of data within a nation's geographic boundaries) with data sovereignty, although this is only one technique among many. Data localization is complicated and may not be well-suited to very broad categories of data. For example, the province of British Columbia at one time prohibited (at least in some circumstances) flows of personal health data outside the country. However, these restrictions proved complex and challenging, and were repealed during the COVID pandemic in part because of the strain it put on the ability to deliver

cloud sovereignty", Policy Options, 29 October 2025, https://policyoptions.irpp.org/2025/10/cloud-sovereignty/. See also: John Ruffalo, "A Blueprint for Canada's Digital Sovereignty", 18 February 2026, Build Canada, https://www.buildcanada.com/memos/digital-sovereignty.

78 H.R.4943 — 115th Congress (2017-2018).

79 See: Barry Appleton, "Whose Law Governs Canadian Data? The CLOUD Act, Executive Agreements and Digital Sovereignty", Balsillie Papers, Special Report, 11 March 2026, https://balsilliepapers.ca/canadian-data/us-cloud-act/.

80 See, e.g. Canadian Centre for Cyber Security, "Using encryptions to keep your sensitive data secure – ITSAP40.016, May 2021, https://www.cyber.gc.ca/en/guidance/using-encryption-keep-your-sensitive-data-secure-itsap40016; Government of Canada White paper: Data Sovereignty and the Public Cloud, 23 April 2026, https://www.canada.ca/en/government/system/digital-government/digital-government-innovations/cloud-services/digital-sovereignty/gc-white-paper-data-sovereignty-public-cloud.html.

virtual health care.[81] Nova Scotia has also enacted a law protecting personal information in the hands of public sector entities against offshore storage and processing without consent or specific provision.[82]

Originally, data sovereignty concerns were linked to control by governments over public sector data, including from the publicly funded health care system. More recently, there have been concerns regarding the reach of foreign states into the potentially vast personal data stores of private corporations. Although the EU has long been concerned about the protection of personal data of EU residents in cross-border data flows, Canada's federal government has largely resisted the enhancement of cross-border data sharing protections in its private sector data protection law in favour of facilitating the flow of data.[83] By contrast, in 2021 reforms to its private sector data protection law, Quebec introduced amendments imposing additional safeguards on companies and public sector entities planning to transfer data outside of the province.[84] It remains to be seen whether more contemporary data sovereignty concerns will push the federal government to require enhanced protections for cross border transfers of private sector data.

Data sovereignty issues that receive less attention but that are also important relate to the ability to control or influence the collection and use of data – especially data of central importance to the state.[85] For example, if a government became in-

---

[81] British Columbia, Ministry of Citizens' Services, "Data Residency Changes", January 2022, https://www2.gov.bc.ca/assets/gov/british-columbians-our-governments/services-policies-for-government/information-management-technology/information-privacy/resources/2021-amendments/foippa_amendments_data_residency.pdf.

[82] Personal Information International Disclosure Protection Act, https://www.canlii.org/en/ns/laws/stat/sns-2006-c-3/96537/sns-2006-c-3.html.

[83] Canada's *Personal Information and Protection of Privacy Act*, SC 2000, c. 5, imposes no special requirements on organizations that transfer data offshore for processing (other than holding them to account). In Bill C-27, supra note 12, its attempt to modernize PIPEDA (which died on the order paper in 2025) the government did not propose any changes to the law to strengthen protection for offshore data transfers.

[84] *An Act to modernize legislative provisions as regards the protection of personal information*, SQ 2021, c 25, s. 27, s. 111.

[85] See, e.g., Peter Johnson & Teresa Scassa, "Who owns the map? Data sovereignty and government spatial data collection, use, and dissemination", Transactions in GIS, 27, 275–289.

creasingly dependent on private sector companies to provide satellite, geospatial, or other core data that it used to collect itself, vulnerabilities are created if these data sources suddenly become unavailable – or if the data is manipulated in any way. In the case of AI, systems trained on data collected in other countries may not be sufficiently representative of Canadian populations or circumstances, leading to poorer performance of these systems in Canada.[86] In the case of generative AI tools, there are concerns about a gradual cultural homogenization or shift towards a dominant foreign culture.[87] Such issues have prompted several European countries to begin developing their own large language or foundation models.[88] These data sovereignty issues are similar to those raised by the Indigenous Data Sovereignty movement. In Canada, the FNIGC's OCAP principles highlight the importance of a holistic approach to data sovereignty which considers not just where data are stored and who controls them, but also who is involved in determining what data will be collected and how they will be used.

3.3 Policy

The ability of a country to set and enforce its own digital policy course is a third

https://doi.org/10.1111/tgis.13024.

86 See, e.g., Canada, Pan-Canadian AI for Health (AI4H) Guiding Principles, 30 January 2025, https://www.canada.ca/en/health-canada/corporate/transparency/health-agreements/pan-canadian-ai-guiding-principles.html.

87 UNESCO, AI & Culture, supra note67; Hamilton Mann, "AI Homogenization Is Shaping the World", Forbes, 1 November 2024, https://www.forbes.com/sites/hamiltonmann/2024/03/05/the-ai-homogenization-is-shaping-the-world/; Alwin de Rooij & Michael Mose Biskjaer, "Does Generative AI Make Us Think Alike? A Systematic Review and Meta-Analysis of Homogenization Effects in Human-AI Co-Creation", 14 April 2026, Tilburg University, https://research.tilburguniversity.edu/files/130048666/Homogenization_in_Human-AI_co-creativity_v1.pdf.

88 See: Anna Desmarais, "Which European countries are building their own sovereign AI to compete in the tech race?", EuroNews, 1 December 2025, https://www.euronews.com/next/2025/12/01/which-european-countries-are-building-their-own-sovereign-ai-to-compete-in-the-tech-race; Louis Westendarp, "Lost in translation: Europe's quest to speak AI", Politico, 19 July 2022, https://www.politico.eu/article/translation-europe-quest-speak-ai-technology/.

dimension of digital sovereignty. Sometimes charting one's own path requires compromise. For example, as a 'middle power', Canada lacks the economic clout of the EU or the US and often finds itself caught between these two major trading partners when it comes to digital policy. The EU's human rights-oriented approach to digital policy seems more aligned with Canadian values and the importance of trade with the EU requires some level of regulatory conformity. At the same time, geographical proximity to the US and trade dependencies make it impossible to ignore the US policy context. Canadians and their businesses can easily move south of the border to find a more hospitable regulatory environment, and this has long been a challenge for Canada.[89] Further, Canadian companies that remain in Canada may struggle when scaling to the US market if their products or services are designed to higher privacy norms or other regulatory requirements.[90]

In other instances, policy sovereignty may be more directly challenged.[91] Canadian attempts at digital policymaking have received a rough reception either from large US companies, from the US government, or both. Such attempts include a digital tax, which was passed into law, but never enforced following strenuous objection from the US.[92] A law designed to compensate Canadian online news services for their content distributed through social media platforms also struggled after Meta refused to comply and instead cut Canadian news content from its services.[93] Canada's new

---

89 See, e.g.: Jonah Prousky, "Canada must tackle its tech talent 'brain drain' problem", Canadian Affairs, 18 July 2024, https://www.canadianaffairs.news/2024/07/18/canada-must-tackle-its-tech-talent-brain-drain-problem/.

90 See, e.g.: Aftab Ahmed, "When Canadian regulations favour corporate concentration", 15 July 2024, Policy Options, https://policyoptions.irpp.org/2024/07/regulations-favour-corporate-concentration/; Business Council of Canada, "Stifled by red tape", 5 September 2025, https://www.thebusinesscouncil.ca/report/stifled-by-red-tape/.

91 See, e.g. Natasha Tusikov, "Canada must ensure its digital sovereignty in the face of U.S. threats", Policy Options 8 May 2025, https://policyoptions.irpp.org/2025/05/digital-sovereignty/.

92 See: White House, "Defending American Companies and Innovators From Overseas Extortion and Unfair Fines and Penalties", 21 February 2025, https://www.whitehouse.gov/presidential-actions/2025/02/defending-american-companies-and-innovators-from-overseas-extortion-and-unfair-fines-and-penalties/.

93 Instead of pulling Canadian news from its services, Google negotiated a deal to pay a fixed sum for the use of Canadian news. See: CBC News, "Google sends $100M in funds to journalism

Online Streaming Act, which was meant to ensure access to, promotion of, and funding for Canadian content through broadcasting and online streaming services has run into strong opposition from US based streaming services.[94]

The 'middle power' status of Canada means that on its own, some digital policy approaches may be beyond its grasp, as it is a small enough market that retaliation or exclusion through geofencing of services remain risks of any policy action. This is where alliances with other similarly situated countries could build a common and strong voice on key issues.

4. AI REGULATION IN CANADA

In June 2022, the federal government introduced its *Artificial Intelligence and Data Act* (AIDA) which was part of Bill C-27.[95] The proposed new law came as a surprise to many as there had been no prior public consultation.[96] The AIDA out-

collective in exchange for Online News Act exemption", 3 January 2025, https://www.cbc.ca/news/entertainment/google-online-news-act-exemption-1.7422690.

94 See, eg.: National Foreign Trade Council (US), "Canada's Streaming Policy is Unworkable", 22 May 2026, https://www.nftc.org/canadas-streaming-policy-is-unworkable/; US Chamber of Commerce, "U.S. Chamber Strongly Condemns Canada's Online Streaming Act", 23 May 2026, https://www.uschamber.com/international/u-s-chamber-strongly-condemns-canadas-online-streaming-act.

95 Bill C-27, supra note12.

96 The lack of proper consultation and public engagement around AI issues has become a recurrent theme in Canada. Lack of consultation prior to the introduction of AIDA has been criticized (see, e.g.: The Dais and Centre for Media, Technology and Democracy (2023) Submission on the proposed Artificial Intelligence and Data Act, https://dais.ca/wp-content/uploads/2023/12/Submission-on-the-Proposed-Artificial-Intelligence-and-Data-Act.pdf) as has the approach of the government to consultation on its next AI Strategy. (See, e.g.: CUPE, "CUPE criticizes federal AI consultation process", (n.d.), https://cupe.ca/cupe-criticizes-federal-ai-consultation-process; British Columbia Civil Liberties Assocation, "Open Letter to the Minister of Artificial Intelligence and Digital Innovation from civil society organizations and individuals opposing "National Sprint" consultation on AI strategy", 31 October 2025, https://bccla.org/policy-submission/open-letter-to-the-minister-of-artificial-intelligence-and-digital-innovation-from-civil-society-organizations-and-individuals-opposing-national-sprint-consultation-on-ai-strategy/. Concerns over the lack of appropriate consultation on the AI Strategy led to the launch of the People's Consultation on AI, https://www.peoplesaiconsultation.ca/.

lined a risk regulation framework for AI governance. Although there were some similarities to the EU's *AI Act*, which was in draft form at the time, there were also significant differences. Like the EU *AI Act*, the AIDA envisaged a sliding scale of obligations for different actors in the AI value chain, although under AIDA, obligations were reserved for "high-impact" as opposed to "high-risk" systems. These included a requirement to evaluate potential risks/harms and to mitigate them. Risks included discriminatory bias.[97] Organizations subject to the law would be required to document their efforts. Oversight would be provided by the Minister of Innovation, Science and Economic Development (ISED), who could require an audit to be conducted, or who, in exceptional circumstances, could take other enforcement actions including imposing administrative monetary penalties. Nonetheless, in notable contrast to the EU AI Act, the approach of the government to regulation under the AIDA was meant to be deliberately 'light-touch' with clear signals being sent to this effect both in the oversight provisions in the legislation and in other communication.[98]

The AIDA left most of its substantive obligations to be developed in regulations (which was a point of criticism of the law).[99] This included the definition of a high-impact system, the steps or measures required to be taken to comply with obligations in the law, and the criteria for imposing AMPs. Criticisms of this and other features of the law led both to a "companion document"[100] which sought to articulate the reasons for the choices made in the proposed law, as well as a raft of proposed amendments to the law from the Minister of ISED.[101] Some of these amendments provided

---

97 AIDA, *supra* note 12, s. 8.

98 Hung et al, *supra* note 7 at 264.

99 See: Teresa Scassa, "Regulating AI in Canada: A Critical Look at the Proposed Artificial Intelligence and Data Act", (2023) 101 Canadian Bar Review 1-30, https://cbr.cba.org/index.php/cbr/article/view/4817.

100 Innovation, Science and Economic Development Canada, The Artificial Intelligence and Data Act (AIDA) – Companion document, February 2023, https://ised-isde.canada.ca/site/innovation-better-canada/en/artificial-intelligence-and-data-act-aida-companion-document.

101 Standing Committee on Industry and Technology, "Correspondence from the Honourable François-Philippe Champagne, Minister of Innovation, Science and Industry - Amendments to AIDA - 2023-11-28", https://www.ourcommons.ca/content/Committee/441/INDU/WebDoc/WD12751351/12751351/MinisterOfInnovationScienceAndIndustry-2023-11-28-Com-

key definitions, greater clarity, and more distance between the Minister and oversight of the law. Others were aimed at adapting the law to generative AI systems, which had been released to the public a few months after the introduction of the bill.

The AIDA reflected both some of the realities of the Canadian context and some of the government's preoccupations at the time. In providing its rationale for the law, the government emphasized the particular nature of Canada's AI industry, which is dominated by small and medium sized enterprises, which might founder if over-regulated.[102] The government also explained its reluctance to provide too much detail and definition in the law because of the rapidly evolving nature of AI technologies.[103] At the same time, the desire to move in the same direction as its closest trading partners was also a motivating factor for the bill.[104] The AIDA looked enough like the EU *AI Act* to create space for potential recognition of Canadian-developed AI systems as being in conformity with EU norms, while at the same time it was meant to be less prescriptive so that its obligations would not create barriers to trade with the US, or send Canadian innovators south of the border.

Canada's constitutional division of powers also impacted the scope of the law, as it relied for its justification on the federal international and interprovincial trade branch of the trade and commerce power. This meant that the law, had it passed, would only have applied to those AI systems planned for commercial deployment internationally or interprovincially.[105] Fully intra-provincial systems would not have been covered, nor would public sector AI (except to the extent that it involved the adoption of systems developed in the private sector). Because the bill was housed with the Minister of ISED, it would also not have supplanted or replaced other AI regulation taking place within other ministries, including health (medical device regulation) or transportation (automated vehicle regulation), to provide two examples. As a result, although the AIDA appeared to be a broad comprehensive risk regulation statute along the lines of the EU AI Act, it was not – either in substance or

bined-e.pdf.

[102] AIDA Companion document, *supra* note 104.

[103] *Ibid*.

[104] *Ibid*.

[105] AIDA, supra note 12, s. 4.

in scope.

Canada's AIDA died on the order paper with the prorogation of Parliament early in 2025. It has not been reintroduced, and the government has clearly signalled that it has no plans to do so.[106] This likely has much to do with the about-face in the United States when it comes to AI regulation. Even a 'light' version of EU-style AI regulation might create too much of a gulf in approach between Canada and the US.

These circumstances create something of a dilemma for Canada. Polling shows that trust in AI in Canada is at a rate that is quite low compared to its peers.[107] This comes at a time when the government is investing heavily in AI and is pushing the adoption and use of AI in the public sector as a way of building trust and supporting the AI industry. The failure of the AIDA has been equated by many with an absence of AI regulation,[108] and this perceived absence of regulation may be a contributing factor to the lack of trust.

The current AI regulatory landscape in Canada has several features. At the federal level it no longer has as a centrepiece a single statute that aims to set the parameters for a broad-based risk-regulation approach. Instead, the government is relying upon multiple modes of regulation, many of them featuring 'soft law' approaches.

4.1 Federal law/policy

4.1.1 Federal Public Sector

The federal government's AI Strategy for the Federal Public Service seeks to encourage adoption of AI to improve service delivery.[109] The Strategy adopts a set of

---

[106] Anindya Sen, Amanda Clarke, Juan Morales, Jonathan Penney and Teresa Scassa, "AI Innovation, Safety and Regulation in Canada: Perspectives from the Editorial Committee", in Special Issue on Artificial Intelligence Safety and Public Policy in Canada, (2026) 52:S1 Canadian Public Policy 1, at 5, https://utppublishing.com/doi/10.3138/cpp.2025-066.

[107] KPMG, Trust, attitudes and use of Artificial Intelligence: A global study 2025, at 28, https://kpmg.com/xx/en/our-insights/ai-and-technology/trust-attitudes-and-use-of-ai.html.

[108] A 2025 KPMG study on trust in AI found that 92% of Canadians are "unaware of any existing, laws, regulations, or policies that govern AI in Canada", *ibid*. at 3.

[109] Treasury Board Secretariat, AI Strategy for the Federal Public Service 2025-2027, 25 February

principles to guide adoption. It also sets as a priority the development of the necessary governance in both law and policy to shape responsible adoption. To this end, it prioritizes "common governance and risk management frameworks for the AI lifecycle".[110] This will likely primarily be achieved through the development of internal policies and Directives. The strategy also indicates that procurement policies will be updated, and that a revision of existing policies and legislation will be carried out.

The current framework for governing AI adoption and use in the public sector principally includes a series of Directives and policy documents issued by the Treasury Board Secretariat. Many of these relate to privacy and data governance more generally and continue to have relevance, although updates and revisions may be required.[111] The Treasury Board *Directive on Automated Decision Making*[112] is one of the few Directives specifically designed for the AI context, and it applies to automated decision-making systems in a cross-section of the federal public sector. The DADM is accompanied by an *Algorithmic Impact Assessment Tool* (AIA).[113]

Procurement can be a powerful tool for AI governance, as the standards set by a potentially major client such as government can shape how an organization develops AI systems. In the case of the federal government, its AI-related procurement approach seems modest at best. It created an AI Source list in 2019 to support AI procurement in the public sector by identifying pre-qualified vendors. This list continues to be updated and now features over 145 suppliers.[114] Suppliers are required to indicate their commitment "to seeking confirmation from our Government of Canada

2025, https://www.canada.ca/en/government/system/digital-government/digital-government-innovations/responsible-use-ai/gc-ai-strategy-overview.html.

[110] *Ibid*.

[111] See, e.g.: Treasury Board Secretariat, Directive on Privacy Practices, https://www.tbs-sct.canada.ca/pol/doc-eng.aspx?id=18309; Treasury Board Secretariat, Policy on Privacy Protection, https://www.tbs-sct.canada.ca/pol/doc-eng.aspx?id=12510.

[112] *Supra* note 10.

[113] Government of Canada, Algorithmic Impact Assessment tool (n.d.), https://www.canada.ca/en/government/system/digital-government/digital-government-innovations/responsible-use-ai/algorithmic-impact-assessment.html.

[114] Government of Canada, Artificial intelligence source list (n.d.), https://www.canada.ca/en/public-services-procurement/services/acquisitions/better-buying/simplifying-procurement-process/artificial-intelligence-source-list.html.

clients of their AIA results in order to best serve Canadians through the application of appropriate mitigation strategies on AI-related products."[115]

Guidance is a soft law tool used in the public sector, particularly where technologies are rapidly evolving. For example, the government has produced a Guide on the Use of Generative AI for the public sector.[116] The Guide suggests best practices, raises issues of potential concerns, and directs users towards relevant laws or Treasury Board policies and directives. They have also recently issued a Guide on the Use of Agentic Artificial Intelligence.[117]

The AI Strategy for the public sector emphasizes the importance of transparency in building trust in AI. One notable commitment is the establishment of a public AI Register. A preliminary version of the Register was shared publicly in November 2025,[118] with a public consultation seeking feedback on the content and structure of the register.[119]

The federal *Privacy Act,* which applies to the federal public sector, and which is overseen by the Privacy Commissioner of Canada, currently governs the protection of personal data. However, the law is sorely out of date. Treasury Board Secretariat has launched a consultation on reform of the *Privacy Act*, which is open until July 10, 2026.[120] The consultation document proposes amendments that would elevate cer-

---

[115] Government of Canada, List of interested Artificial Intelligence (AI) suppliers, (n.d.)https://www.canada.ca/en/government/system/digital-government/digital-government-innovations/responsible-use-ai/list-interested-artificial-intelligence-ai-suppliers.html. Note that the commitment does not create any legally binding obligations.

[116] Government of Canada, Guide on the use of generative artificial intelligence, 3 June 2025, https://www.canada.ca/en/government/system/digital-government/digital-government-innovations/responsible-use-ai/guide-use-generative-ai.html.

[117] Government of Canada, Guide on the Use of Agentic Artificial Intelligence, 22 May 2026, https://www.canada.ca/en/government/system/digital-government/digital-government-innovations/responsible-use-ai/guide-use-agentic-artificial-antelligence.html.

[118] Government of Canada, AI Register (Minimum Viable Product), (n.d.) https://open.canada.ca/data/en/dataset/fcbc0200-79ba-4fa4-94a6-00e32facea6b.

[119] Government of Canada, "Canada launches first register of AI uses in federal government", 28 November 2025, https://www.canada.ca/en/treasury-board-secretariat/news/2025/11/canada-launches-first-register-of-ai-uses-in-federal-government.html.

[120] Treasury Board Secretariat, Review of Privacy Act, *supra* note 52.

tain aspects of the *Directive on Automated Decision Making* into the legislation, and that would facilitate horizontal sharing of personal data across government to facilitate better use of personal data in public sector AI.[121] Other modernization recommendations relevant in the AI context relate to improving accountability through stronger enforcement.

4.1.2 Private Sector

The federal government may have abandoned risk-regulation legislation for the private sector, but it has embraced soft-law risk regulation approaches. One of these is its Voluntary Code of Conduct on the Responsible Development of AI[122] which sets out a series of measures that organizations seeking to comply with the Code must implement. It also allows organizations that consider themselves to be compliant to indicate their voluntary compliance by signing on. At the time of writing, forty-six companies had signed the Code. These consist of both large and small corporations, as well as Canadian and offshore companies.

Standards are another way to encourage voluntary compliance with best practices for artificial intelligence development and deployment. Canada is involved in both domestic and international standards development.[123] In Canada, the Digital Governance Standards Institute has been accredited to develop standards for digital and data governance, and its work has included standards relating to artificial intelligence, including ethical AI development for small and medium sized enterprises, and ethical AI in clinical decision support.[124] Canada hopes to leverage the Canada Eu-

---

[121] *Ibid*.

[122] Government of Canada, Voluntary Code of Conduct on the Responsible Development and Management of Advanced Generative AI Systems, September 2023, https://ised-isde.canada.ca/site/ised/en/voluntary-code-conduct-responsible-development-and-management-advanced-generative-ai-systems

[123] Standards Council of Canada, Canada contributes to global collaboration on responsible AI at the International AI Standards Summit 2025, 16 December 2025, https://scc-ccn.ca/resources/news/canada-contributes-global-collaboration-responsible-ai-international-ai-standards.

[124] Digital Governance Council, Work Program: Period: April 2, 2026 to May 1, 2026, https://dgc-cgn.org/standards/work-program/.

rope Trade Agreement as a means of identifying key standards and agreeing on mutual recognition of conformity assessments as a means of facilitating trade in AI goods and services.[125]

Proposed amendments to the federal *Red Tape Reduction Act*[126] are meant to facilitate the creation of regulatory sandboxes at the federal level, although it remains unclear to what extent these sandboxes will be directed towards AI innovation.[127] Canada is also in discussions with the EU over the potential for Canadian companies to participate in EU AI regulatory sandboxes as a means of gaining access to EU markets.[128]

Existing regulators are also taking on some of the work of AI governance at the federal level. The Privacy Commissioner of Canada has jurisdiction over private sector data protection at a national level under the *Personal Information Protection and Electronic Documents Act* (PIPEDA).[129] This statute is badly in need of reform, and predates artificial intelligence technologies at the commercial scale. Nevertheless, it has been interpreted and applied to contexts involving data scraping for both facial recognition technologies[130] and the development of Large Language Models. [131] The federal Commissioner, in conjunction with provincial and territorial commissioners, has also issued *Principles for responsible, trustworthy and privacy protec-*

---

[125] Global Affairs Canada, Report: Comprehensive Economic and Trade Agreement (CETA) 7th meeting of the Committee on Trade in Goods, October 15-16, 2025, https://www.international.gc.ca/trade-commerce/trade-agreements-accords-commerciaux/agr-acc/ceta-aecg/2026-02-16-report-rapport.aspx?lang=eng

[126] Red Tape Reduction Act, SC 2015, c 12,

[127] *Budget Implementation Act*, 2025, No. 1, SC 2025, c. 3, ss. 203-204.

[128] Global Affairs Canada, *supra* note 129 at 5.4.2.

[129] S.C. 2000, c. 5.

[130] See: Office of the Privacy Commissioner of Canada, Joint investigation of Clearview AI, Inc. by the Office of the Privacy Commissioner of Canada, the Commission d'accès à l'information du Québec, the Information and Privacy Commissioner for British Columbia, and the Information Privacy Commissioner of Alberta, 2 February 2021, https://www.priv.gc.ca/en/opc-actions-and-decisions/investigations/investigations-into-businesses/2021/pipeda-2021-001/.

[131] See: Office of the Privacy Commissioner of Canada, Joint Investigation of OpenAI OpCo, LLC, 6 May 2026, https://www.priv.gc.ca/en/opc-actions-and-decisions/investigations/investigations-into-businesses/2026/pipeda-2026-002/.

tive generative AI technologies.[132] Anticipated amendments to PIPEDA include a right to an explanation of automated decision-making.[133]

Recognizing the importance of regulatory collaboration to address emerging issues, Canada has established the Canadian Digital Regulator's Forum, which consists of The Canadian Radio-television and Telecommunications Commission, the Competition Bureau, the Office of the Privacy Commissioner of Canada and the Copyright Board of Canada. The objective behind the formation of this group was to enhance cooperation and collaboration on digital regulatory issues.

The Competition Bureau of Canada held a public consultation in 2024[134] on AI in competition to assist it in developing its approaches towards AI technologies and competition.[135] More recently it issued a discussion paper and consultation on algorithmic pricing and competition.[136] Health Canada regulates medical devices, and has recently issued Pre-market guidance for machine learning-enabled medical devices that are subject to the approval process.[137] The government has also consulted on AI in relation to copyright law, although no law reform has yet emerged from this process.[138]

---

[132] Office of the Privacy Commissioner of Canada, Principles for responsible, trustworthy and privacy-protective generative AI technologies, 7 December 2023, https://www.priv.gc.ca/en/privacy-topics/technology/artificial-intelligence/gd_principles_ai/.

[133] See, e.g., Bill C-27, *supra* note 12, s. 63(3).

[134] Competition Bureau of Canada, Artificial intelligence and competition: Discussion Paper, March 2024, https://competition-bureau.canada.ca/en/how-we-foster-competition/education-and-outreach/artificial-intelligence-and-competition.

[135] Competition Bureau of Canada, Consultation on Artificial Intelligence and Competition: What We Heard, 27 January 2025, https://competition-bureau.canada.ca/en/how-we-foster-competition/education-and-outreach/consultation-artificial-intelligence-and-competition-what-we-heard.

[136] Competition Bureau of Canada, Algorithmic pricing and competition: Discussion paper, 10 June 2025, https://competition-bureau.canada.ca/en/how-we-foster-competition/education-and-outreach/publications/algorithmic-pricing-and-competition-discussion-paper.

[137] Health Canada, Pre-market guidance for machine learning-enabled medical devices, 1 April 2026, https://www.canada.ca/en/health-canada/services/drugs-health-products/medical-devices/application-information/guidance-documents/pre-market-guidance-machine-learning-enabled-medical-devices.html.

[138] For the original consultation document, see: Innovation, Science and Economic Development,

This activity by different regulators suggests that a sectoral approach to AI regulation is evolving. However, there may be a need for a more centralized agency to play a coordinating and convening role across the federal government and in seeking cooperation and harmonized approaches from the provinces.

4.2 Provincial law/policy

Canadian provinces and territories have jurisdiction over aspects of artificial intelligence, although there has been relatively little action to regulate AI by provincial governments. Ontario is the first and only province to specifically legislate for AI governance, although it has done so in a statute that creates a framework that requires regulations to provide substance to the scheme. At the time of writing, no regulations are in the pipeline. Ontario also has issued a Responsible Use of Artificial Intelligence Directive, which applies to Ontario ministries and provincial agencies. [139] Manitoba has introduced an AI regulation bill which largely mirrors the Ontario bill, and which will therefore also require regulations to take effect. [140] In Quebec, the Ministère de la Cybersécurité et du Numérique has adopted a Statement of Principles for the Responsible Use of Artificial Intelligence by Public Bodies. [141] Other provinces have issued more general guidelines for responsible public sector AI use. [142]

Canada, "Consultation on Copyright in the Age of Generative Artificial Intelligence", 5 December 2024, https://ised-isde.canada.ca/site/strategic-policy-sector/en/marketplace-framework-policy/consultation-paper-consultation-copyright-age-generative-artificial-intelligence; for the report on the consultation, see: Innovation, Science and Economic Development, Canada, Consultation on Copyright in the Age of Generative Artificial Intelligence: What we heard report, 11 February 2025, https://ised-isde.canada.ca/site/strategic-policy-sector/en/marketplace-framework-policy/consultation-copyright-age-generative-artificial-intelligence-what-we-heard-report.

139 Ontario, Responsible Use of Artificial Intelligence Directive,7 January 2022, https://www.ontario.ca/page/responsible-use-artificial-intelligence-directive#section-2.

140 Manitoba, Bill 51, The Public Sector Artificial Intelligence and Cybersecurity Governance Act, 3rd Sess. 43rd Legislature, https://web2.gov.mb.ca/bills/43-3/b051e.php.

141 Arrêté numéro 2024-02 du ministre de la Cybersécurité et du Numérique en date du 27 juin 2024, https://www.publicationsduquebec.gouv.qc.ca/fileadmin/gazette/pdf_encrypte/lois_reglements/2024F/83874.pdf

142. See, e.g., British Columbia, "Policy on the use of generative AI", 9 May 2025, https://dig-

Some provinces have moved forward with new legislation (or with amendments to existing legislation) to address specific AI-related issues. For example, Manitoba has introduced a bill to amend its Business Practices Act to prohibit algorithmic price discrimination.[143] Saskatchewan, Manitoba, British Columbia and Quebec have amended their legislation regarding non-consensual intimate images to specifically include deepfakes, and Alberta has signalled its intention to do the same.[144] Alberta has also amended its law relating to elections to specifically address deepfakes used in the elections context.[145] Ontario has enacted a provision in its employment standards legislation requiring transparency with respect to the use of AI in hiring practices.[146]

Several provinces have also updated their public sector data protection laws to enable better horizontal data sharing across government, and to address other AI related issues. For example, Quebec has added a right to an explanation to its public sector data protection law.[147] Alberta has added a requirement to give notice of the intent to collect information for the purposes of automated decision making, as well as requirements regarding accuracy and data retention in that context.[148]

Existing regulators may also apply their laws to issues that arise in relation to the use of artificial intelligence. Both the Quebec and the Ontario Information and Privacy Commissioners have addressed artificial intelligence-related issues in some inves-

ital.gov.bc.ca/policies-standards/generative-ai-policy/; Saskatchewan, "Guidelines for Government of Saskatchewan Employees in Using Generative Artificial Intelligence", (n.d.) https://taskroom.saskatchewan.ca/-/media/project/taskroom/documents/guide/artificial-intelligence-guidelines.pdf; Alberta, Artificial Intelligence Usage Policy, 1 May 2025, https://manuals.alberta.ca/imt-policy-instruments-portal/policies/cybersecurity/artificial-intelligence-usage-policy/. .

143 Manitoba, Bill 49, *The Business Practices Amendment Act*, 3rd Sess, 43rd Legislature, https://web2.gov.mb.ca/bills/43-3/b049e.php.

144 Jack Farrell, "Alberta to change legislation allowing lawsuits over sharing deepfake intimate images", CBC News, 24 March 2026, https://www.cbc.ca/news/canada/calgary/albreta-deepfake-legislation-9.7139630.

145 *Election Finances and Contributions Disclosure Act*, RSA 2000, c E-2, s. 44.9499995.

146 *Working for Workers Four Act*, 2024, SO 2024, c 3, s. 8.4.

147 *Act respecting Access to documents held by public bodies and the Protection of personal information*, CQLR c A-2.1, s. 65.2.

148 *Protection of Privacy Act*, SA 2024, c P-28.5, s. 5.1, s. 6.

tigations,[149] and both BC and Quebec have issued decisions regarding the use of facial recognition technologies in the retail sector.[150] In addition to investigating AI issues that fall within their jurisdiction, provincial privacy regulators have been involved in issuing guidance on AI and privacy related issues. Four commissioners have, for example, issued guidance on the use of AI medical scribes.[151] Provincial information and privacy commissioners have joined with the federal Privacy Commissioner in issuing guidance on the use of facial recognition technologies and on generative AI and privacy.[152]

Provincial human rights commissioners may also have jurisdiction over some

---

149 See, e.g.: Ontario Information and Privacy Commissioner, Privacy Complaint Report PI21-00001, 28 February 2024, https://norma.lexum.com/ipc-cipvp/privacy/en/item/521580/index.do; Ontario Information and Privacy Commissioner, Privacy Complaint Report PX24-00001, 11 June 2025, https://norma.lexum.com/ipc-cipvp/privacy/en/item/521985/index.do; Commission d'accès à l'information du Quebec, Centre de services scolaire du Val-des-Cerfs (anciennement Commission scolaire du Val-des-Cerfs, 9 November 2022, https://www.cai.gouv.qc.ca/uploads/pdfs/decisions-en-surveillance/1020040-S_decision-cai-centre-services-scolaire-val-cerfs-2022-11-09.pdf?v=1764365041.

150 See, e.g., British Columbia Office of the Information and Privacy Commissioner, Investigation Report 23-02, Canadian Tire Associate Dealers' use of facial recognition technology, April 2023, https://www.oipc.bc.ca/documents/investigation-reports/2618; Commission d'accès à l'information du Québec, Enquête à l'égard de l'Auberge du lac Sacacomie inc., C.A.I. 1014137-S, 7 April 2022; Enquête à l'égard d'Imprimeries Transcontinental inc., C.A.I. 1024350-S, 4 September 2024.

151 See, e.g: Ontario Information and Privacy Commissioner, AI Scribes: Key Considerations for the Health Sector, 28 January 2026, https://www.ipc.on.ca/en/resources/ai-scribes-key-considerations-health-sector; British Columbia Information and Privacy Commissioner, PIPA and AI Scribes: best practices for healthcare organizations in BC, January 2026, https://www.oipc.bc.ca/documents/guidance-documents/3082; Saskatchewan Information and Privacy Commissioner , Checklist for Healthcare Organizations Considering the Use of an AI Scribe, 8 September 2025, https://oipc.sk.ca/assets/checklist-for-healthcare-organizations-considering-the-use-of-an-AI-scribe.pdf; Alberta Information and Privacy Commissioner, Artificial Intelligence (AI) Scribe Privacy Impact Assessment Guidance, September 2025, https://oipc.ab.ca/office-of-the-information-and-privacy-commissioner-issues-guidance-on-ai-scribe-tools-for-custodians-under-health-information-act/.

152 Office of the Privacy Commissioner of Canada, Privacy guidance on facial recognition for police agencies, 2 May 2022, https://www.priv.gc.ca/en/privacy-topics/surveillance/police-and-public-safety/gd_fr_202205/; Office of the Privacy Commissioner of Canada, Principles, *supra* note 132.

matters involving discriminatory uses of AI technologies. The Ontario Human Rights Commission has issued joint AI principles with the Ontario Office of the Information and Privacy Commissioner,[153] and it has also joined with the Law Commission of Ontario to produce a voluntary AI Human Rights Impact Assessment tool.[154]

There are numerous other regulatory bodies at the provincial level that are involved in some form of AI regulation, most typically at the level of guidance, principles or policy. This is the case with provincial law societies (which govern the practice of law), as well as other bodies that govern regulated professions. Some provincial law societies have created AI regulatory sandboxes in order to experiment with the use of AI technologies to improve access to justice in safe and effective ways.[155]

The law relating to civil liability – another area of importance to AI regulation – falls primarily under provincial jurisdiction in Canada. This means that the creation of statutory torts is within the competence of provincial legislatures, while the common law in nine of Canada's ten provinces evolves according to case law in each of those jurisdictions. Quebec, which maintains a civil law system for its private law, governs civil liability principally through its *Civil Code of Quebec*.[156] As a result, issues such as civil liability for faulty or harmful AI systems are addressed at the provincial level. The federal government, with its criminal law power, can criminalize harmful conduct. This means that for some subject matter – such as deepfakes – both federal and provincial governments may have legislative competence with respect to different dimensions of the problem.

---

153 Ontario Information and Privacy Commissioner and Ontario Human Rights Commission, Principles for the responsible use of artificial intelligence, 21 January 2026, https://www.ipc.on.ca/en/media/6116/download?attachment.

154 Law Commission of Ontario and Ontario Human Rights Commission, Human Rights AI Impact Assessment, November 2024, https://www.ohrc.on.ca/en/human-rights-ai-impact-assessment.

155 See: Law Society of Ontario, Access to Innovation, online: https://lso.ca/about-lso/access-to-innovation/about-a2i; Law Society of British Columbia, Innovation Sandbox, https://www.lawsociety.bc.ca/about-us/priorities/innovation-sandbox/.

156 CQLR c CCQ-1991.

5. CONCLUSION

Canada's 'middle power' approach to AI regulation is increasingly shaped by the realities of geopolitical instability and economic dependence. Canada's earlier ambition to position itself as both an AI innovator and a leader in principles-based AI governance has been complicated by the rise of sovereignty concerns, shifting global alliances, and the growing divergence between European and American approaches to AI regulation. In this context, Canada faces the difficult task of balancing economic competitiveness, innovation, and trade dependencies with the need to protect rights, democratic values, and public trust.

Canada's experience also demonstrates that AI regulation cannot be understood solely as a matter of technical governance. Federalism creates structural constraints on comprehensive centralized regulation, which can slow regulatory responses and created fragmented approaches. Yet federalism is also meant to foster dialogue, experimentation, and collaborative governance across jurisdictions. The persistent need to build alliances and find common ground may be valuable middle power skills. Canada's efforts at reconciliation with its diverse Indigenous Peoples are also lessons in how governance must include meaningful recognition of diverse forms of self-determination. Notably, in the AI context, there are important lessons here regarding data sovereignty and culturally grounded approaches to AI development and deployment. These realities make clear that AI governance in Canada will necessarily be pluralistic, layered, and negotiated.

Canada's retreat from its proposed Artificial Intelligence and Data Act early in 2025 marked a shift in course but does not signal the end of AI governance in Canada. Rather, it suggests a transition toward a more fragmented and hybrid model that relies on sectoral regulation, public sector directives, procurement rules, data governance measures, soft law instruments, and international partnerships. Whether this model will be sufficient to build public trust remains uncertain, particularly given persistent concerns regarding AI harms, labour impacts, democratic integrity, surveillance, bias, and cultural sovereignty.

For middle power countries such as Canada, the challenge is not just to choose between European-style regulation and American market-driven approaches. It is

instead to identify forms of governance that preserve policy autonomy while enabling participation in an increasingly competitive global AI economy. In this respect, Canada's future role may depend less on acting alone than on building coalitions with other middle powers that share commitments to human rights, democratic governance, diversity, and responsible innovation.

REFERENCES

Ahmed, Aftab. "When Canadian regulations favour corporate concentration", 15 July 2024, Policy Options, https://policyoptions.irpp.org/2024/07/regulations-favour-corporate-concentration/.

Alberta, Artificial Intelligence Usage Policy, 1 May 2025, https://manuals.alberta.ca/imt-policy-instruments-portal/policies/cybersecurity/artificial-intelligence-usage-policy/.

Alberta Information and Privacy Commissioner , *Artificial Intelligence (AI) Scribe Privacy Impact Assessment Guidance*, September 2025, https://oipc.ab.ca/office-of-the-information-and-privacy-commissioner-issues-guidance-on-ai-scribe-tools-for-custodians-under-health-information-act/.

Alberta Machine Intelligence Institute, "Pan-Canadian AI Strategy", (n.d.), https://www.amii.ca/about/pan-canadian-ai-strategy.

Alhamad, Rami. "Canada Is Spending Billions on AI. Why Are Companies Still Fleeing?", The Walrus, 20 May 2026, https://thewalrus.ca/canada-is-spending-billions-on-ai-why-are-companies-still-fleeing/.

Appleton, Barry. "Whose Law Governs Canadian Data? The CLOUD Act, Executive Agreements and Digital Sovereignty", Balsillie Papers, Special Report, 11 March 2026, https://balsilliepapers.ca/canadian-data/us-cloud-act/.

Arrêté numéro 2024-02 du ministre de la Cybersécurité et du Numérique en date du 27 juin 2024 , https://www.publicationsduquebec.gouv.qc.ca/fileadmin/gazette/pdf_encrypte/lois_reglements/2024F/83874.pdf.

Benz, Arthur and Sonnicksen, Jared. "Patterns of federal democracy: tensions, friction, or balance between two government dimensions", European Political Science Review (2017), 9:1, 3–25.

British Columbia Information and Privacy Commissioner, *PIPA and AI Scribes: best practices for healthcare organizations in BC*, January 2026, https://www.oipc.bc.ca/documents/guidance-documents/3082.

British Columbia Information and Privacy Commissioner, Investigation Report 23-02, Canadian Tire Associate Dealers' use of facial recognition technology , April 2023, https://www.oipc.bc.ca/documents/investigation-reports/2618.

British Columbia, "Policy on the use of generative AI", 9 May 2025, https://digital.gov.bc.ca/policies-

standards/generative-ai-policy/.

British Columbia, Ministry of Citizens' Services, "Data Residency Changes", January 2022, https://www2.gov.bc.ca/assets/gov/british-columbians-our-governments/services-policies-for-government/information-management-technology/information-privacy/resources/2021-amendments/foippa_amendments_data_residency.pdf.

British Columbia Civil Liberties Assocation , "Open Letter to the Minister of Artificial Intelligence and Digital Innovation from civil society organizations and individuals opposing "National Sprint" consultation on AI strategy ", 31 October 2025, https://bccla.org/policy-submission/open-letter-to-the-minister-of-artificial-intelligence-and-digital-innovation-from-civil-society-organizations-and-individuals-opposing-national-sprint-consultation-on-ai-strategy/.

Burgess, Matt and Elliot, Vittoria. "The EU is Going Through a Trump-Fueled Breakup with Big Tech", WIRED, 21 May 2026, https://www.wired.com/story/the-eu-is-going-through-a-trump-fueled-breakup-with-big-tech/.

Business Council of Canada, "Stifled by red tape", 5 September 2025, https://www.thebusinesscouncil.ca/report/stifled-by-red-tape/.

Butler, Martha and Tiedemann, Marlisa. The Federal Role in Health and Health Care, Library of Parliament, 2011, Publication No. 2011-91-E, https://lop.parl.ca/sites/PublicWebsite/default/en_CA/ResearchPublications/201191E.

Cairns, A.C., "The Judicial Committee and Its Critics" (1971) 4 Canadian Journal of Political Science 301.

Canada, "Canada launches national initiative to build large-scale AI supercomputing capacity ", 15 April 2026, https://www.canada.ca/en/innovation-science-economic-development/news/2026/04/canada-launches-national-initiative-to-build-large-scale-ai-supercomputing-capacity.html.

Canada, Canada's Digital Charter (2019), https://ised-isde.canada.ca/site/innovation-better-canada/en/canadas-digital-charter-trust-digital-world .

Canada, Directive on Automated Decision Making (2019), https://www.tbs-sct.canada.ca/pol/doc-eng.aspx?id=32592&section=html#appA.

Canada, Industry, Science and Economic Development, "Pan-Canadian Artificial Intelligence Strategy", (2017), https://ised-isde.canada.ca/site/ised/en/pan-canadian-artificial-intelligence-strategy.

Canada, Pan-Canadian AI for Health (AI4H) Guiding Principles, 30 January 2025, https://www.canada.ca/en/health-canada/corporate/transparency/health-agreements/pan-canadian-ai-guiding-principles.html.

Canada, Truth and Reconciliation Commission of Canada. Canada's residential schools: the final report of the Truth and Reconciliation Commission of Canada (2015), https://publications.gc.ca/site/eng/9.807830/publication.html.

Canadian Centre for Cyber Security, "Using encryptions to keep your sensitive data secure – ITSAP40.016, May 2021, https://www.cyber.gc.ca/en/guidance/using-encryption-keep-your-sensitive-data-secure-itsap40016.

Carroll Rainie, Stephanie; Kukutai, Tahu; Walter, Maggie; Figueroa-Rodriguez, Oscar Luis; Walker, Jennifer & Axelsson, Per. "Indigenous Data Sovereignty" in The State of Open Data 300, 300–19 (Davies et al. ed., 2019).

Chatterjee, Rhitu. "Their teenage sons died by suicide. Now, they are sounding an alarm about AI chatbots", NPR online, 19 September 2025, https://www.npr.org/sections/shots-health-news/2025/09/19/nx-s1-5545749/ai-chatbots-safety-openai-meta-characterai-teens-suicide.

Competition Bureau of Canada, "Algorithmic pricing and competition: Discussion paper", 10 June 2025, https://competition-bureau.canada.ca/en/how-we-foster-competition/education-and-outreach/publications/algorithmic-pricing-and-competition-discussion-paper.

CBC News, " Google sends $100M in funds to journalism collective in exchange for Online News Act exemption ", 3 January 2025, https://www.cbc.ca/news/entertainment/google-online-news-act-exemption-1.7422690.

Cerutti, Eugenio M.; Garcia Pascual, Antonio I.; Kido, Yosuke; Li, Longji; Melina, Giovanni; Mendes Tavares, Marina; and Wingender, Philippe. "The Global Impact of AI: Mind the Gap", IMF Working Papers 2025, 076 (2025), https://doi.org/10.5089/9798229008570.001 .

Champagne, François-Philippe, "Canada's Artificial Intelligence Revolution: Leading with Purpose ", (2024), https://www.innovatingcanada.ca/technology/future-of-ai-2024/canadas-artificial-intelligence-revolution-leading-with-purpose/.

Chouksey, Apoorva; Rajan, Aneha K.; Gurjar, Vikas; Tiwari, Rajnarayan; Kumar Mishra, Pradyumna. "The green paradox: The climate, environmental, and sustainability implications of artificial intelligence ", (2026) 6 Global Environmental Change Advances 100029, https://doi.org/10.1016/j.gecadv.2025.100029.

CIFAR, "The Pan-Canadian AI Strategy," (n.d.), https://cifar.ca/ai/.

Clement, Andrew. "Canadian Network Sovereignty: A Strategy for Twenty-First-Century National Infrastructure Building", 26 March 2018, CIGI Online, https://www.cigionline.org/articles/canadian-network-sovereignty/.

Commission d'accès à l'information du Québec, Enquête à l'égard de l'Auberge du lac Sacacomie inc. , C.A.I. 1014137-S, 7 April 2022.

Commission d'accès à l'information du Québec, Enquête à l'égard d'Imprimeries Transcontinental inc. , C.A.I. 1024350-S, 4 September 2024.

Commission d'accès à l'information du Quebec, Centre de services scolaire du Val-des-Cerfs (anciennement Commission scolaire du Val-des-Cerfs , 9 November 2022, https://www.cai.gouv.qc.ca/uploads/pdfs/decisions-en-surveillance/1020040-S_decision-cai-

centre-services-scolaire-val-cerfs-2022-11-09.pdf?v=1764365041.

Competition Bureau, Unlocking the power of health data, June 23, 2022, https://competition-bureau.canada.ca/en/unlocking-power-health-data.

Competition Bureau of Canada, Algorithmic pricing and competition: Discussion paper , 10 June 2025, https://competit i on-bureau.canada.ca/en/how-we-foster-competition/education-and-outreach/publications/algorithmic-pricing-and-competition-discussion-paper.

Competition Bureau of Canada, *Artificial intelligence and competition : Discussion Paper*, March 2024, https://compet i tion-bureau.canada.ca/en/how-we-foster-competition/education-and-outreach/artificial-intelligence-and-competition.

Competition Bureau of Canada, Consultation on Artificial Intelligence and Competition: What We Heard , 27 January 2025, https://competition-bureau.canada.ca/en/how-we-foster-competition/education-and-outreach/consultation-artificial-intelligence-and-competition-what-we-heard.

CUPE, "CUPE criticizes federal AI consultation process", (n.d.), https://cupe.ca/cupe-criticizes-federal-ai-consultation-process.

de Rooij, Alwin & Mose Biskjaer, Michael. "Does Generative AI Make Us Think Alike? A Systematic Review and Meta-Analysis of Homogenization Effects in Human-AI Co-Creation", 14 April 2026, Tilburg University, https://research.tilburguniversity.edu/files/130048666/Homogenization_in_Human-AI_co-creativity_v1.pdf.

Desmarais, Anna. "Which European countries are building their own sovereign AI to compete in the tech race? ", EuroNews , 1 December 2025, https://www.euronews.com/next/2025/12/01/which-european-countries-are-building-their-own-sovereign-ai-to-compete-in-the-tech-race.

Digital Governance Council, Work Program : Period: April 2, 2026 to May 1, 2026 , https://dgc-cgn.org/standards / work-program/.

Dorn, Anna. "How AI can adapt to Indigenous knowledge", Medium, 7 November 2025, https://medium.com/blog/how-ai-can-adapt-to-indigenous-knowledge-2ae3cd442096.

Economic Development Research Partners, "Artificial Intelligence Impact on Labor Market s: Literature Review", International Economic Development Council, 13 March 2025, https://www.iedconline.org/clientuploads/EDRP%20Logos/AI_Impact_on_Labor_Markets.pdf.

Farrell, Jack. "Alberta to change legislation allowing lawsuits over sharing deepfake intimate images ", CBC News, 24 March 2026, https://www.cbc.ca/news/canada/calgary/albreta - deepfake-legislation-9.7139630.

Filippucci , F. et al. (2024), "The impact of Artificial Intelligence on productivity, distribution and growth: Key mechanisms, initial evidence and policy challenges", OECD Artificial Intelligence Papers, No. 15, OECD Publishing, Paris, https://doi.org/10.1787/8d900037-en.

First Nations Information Governance Centre, Ownership, Control, Access and Possession (OCAP™)

(2014).

Gammon, Keri. "Pandemics and Pandemonium: Constitutional Jurisdiction Over Public Health", 2006 15 Dalhousie Journal of Legal Studies 1.

Geist, Michael. "The US isn't happy about Canada's quest for digital sovereignty", 9 April 2026, Globe and Mail, https://www.theglobeandmail.com/business/commentary/article-the-us-isnt-happy-about-canadas-quest-for-digital-sovereignty/.

Geist, Michael. "Why the Trump Trade Threats Will Place Canadian Digital, Cultural, and AI Policy Under Pressure", February 23, 2025, https://mgeist.substack.com/p/why-the-trump-trade-threats-will.

Gillespie, N.; Lockey, S.; Ward, T.; Macdade , A.; & Hassed, G. (2025). Trust, attitudes and use of artificial intelligence: A global study 2025. The University of Melbourne and KPMG. DOI 10.26188/28822919.

Global Affairs Canada, Report: Comprehensive Economic and Trade Agreement (CETA) 7th meeting of the Committee on Trade in Goods , October 15-16, 2025, https://www.internati o nal.gc.ca/trade-commerce/trade-agreements-accords-commerciaux/agr-acc/ceta-aecg/2026-02-16-report-rapport.aspx?lang=eng.

Goodfellow Craig, Ashley. " How York U is using technology to protect Indigenous languages", YorkU, 6 November 2025, https://www.yorku.ca/yfile/2025/11/06/how-york-u-is-using-technology-to-protect-indigenous-languages/.

Government of Canada, "Canada moves toward safe and responsible artificial intelligence", 6 March 2025, https://www.ca n ada.ca/en/innovation-science-economic-development/news/2025/03/canada-moves-toward-safe-and-responsible-artificial-intelligence.html.

Government of Canada Cloud Security Risk Management Approach and Procedures, 22 October 2025, https://www.canada.ca/en/government/system/digital-government/digital-government-innovations/cloud-services/cloud-security-risk-management-approach-procedures.html.

Government of Canada Cloud Adoption Strategy, 2018, https://www.canada.ca/en/government/system/digital-government/digital-government-innovations/cloud-services/government-canada-cloud-adoption-strategy.html.

Government of Canada, AI Register (Minimum Viable Product) , (n.d.) https://open.canada.ca/data/en/dataset/fcbc0200-7 9 ba-4fa4-94a6-00e32facea6b.

Government of Canada, Algorithmic Impact Assessment tool (n.d.), https://www.canada.ca/en/gove r nment/system/digital-government/digital-government-innovations/responsible-use-ai/algorithmic-impact-assessment.html.

Government of Canada, Artificial intelligence source list (n.d.), https://www.canada.ca/en/public-services- p rocurement/services/acquisitions/better-buying/simplifying-procurement-process/artificial-intelligence-source-list.html.

Government of Canada, Guide on the Use of Agentic Artificial Intelligence, 22 May 2026, https://www.canada.ca/en/government/system/digital-government/digital-government-innovations/responsible-use-ai/guide-use-agentic-artificial-antelligence.html.

Government of Canada, Guide on the use of generative artificial intelligence , 3 June 2025, https://www.canada.ca/en / government/system/digital-government/digital-government-innovations/responsible-use-ai/guide-use-generative-ai.html.

Government of Canada, List of interested Artificial Intelligence (AI) suppliers , (n.d.) https://www.canada.ca/en/government/system/di g ital-government/digital-government-innovations/responsible-use-ai/list-interested-artificial-intelligence-ai-suppliers.html .

Government of Canada White Paper: Data Sovereignty and the Public Cloud, 23 April 2026, https://www.canada.ca/en/government/system/digital-government/digital-government-innovations/cloud-services/digital-sovereignty/gc-white-paper-data-sovereignty-public-cloud.html.

Government of Canada, Voluntary Code of Conduct on the Responsible Development and Management of Advanced Generative AI Systems , September 2023, https://ised-is d e.canada.ca/site/ised/en/voluntary-code-conduct-responsible-development-and-management-advanced-generative-ai-systems

Government of Canada, "Canada launches first register of AI uses in federal government ", 28 November 2025, https://www . canada.ca/en/treasury-board-secretariat/news/2025/11/canada-launches-first-register-of-ai-uses-in-federal-government.html.

Health Canada, Pre-market guidance for machine learning-enabled medical devices , 1 April 2026, https://www.canada.ca/en/healt h -canada/services/drugs-health-products/medical-devices/application-information/guidance-documents/pre-market-guidance-machine-learning-enabled-medical-devices.html.

Hummel, Patrick; Braun, Matthias; Tretter, Max and Dabrock, Peter. "Data sovereignty: A review", (2021) *Big Data and Society* 1, https://d-nb.info/1282975056/34.

Hung, K.H., Li, L.X., Katrak , M. and Attard-Frost, B. (2025) "Canada's high-stakes artificial intelligence gamble: innovation policy, techno-nationalism and the political economy of jobs" , *Global Political Economy*, 4(2): 256–269, DOI:10.1332/26352257Y2025D000000040.

Iliadis, Andrew; Tripodi, Francesca; Dave, Aashka; Kay Jones, Leslie; Acker, Amelia, and Ford, Heather. "Infrastructures of Manipulation", 2023 AoIR2023, https://spir.aoir.org/ojs/index.php/spir/article/view/13539.

Indigenous Population Health: Our Data, Our Stories , Indigenous Primary Health Care Council, 2025, https://iphcc.ca/wp-content/uploads/ninja-forms/2/D4-Data-Governance-FINAL-1.pdf.

Innovation, Science and Economic Development , Canada, " Consultation on Copyright in the Age of Generative Artificial Intelligence ", 5 December 2024, https://ised-isde.canada.ca/site/strategic-policy-sector/en/marketplace-framework-policy/consultation-paper-consultation-copyright-

age-generative-artificial-intelligence.

Innovation, Science, and Economic Development, "Canadian Sovereign AI Compute Strategy", 15 April 2026, https://ised-isde.canada.ca/site/ised/en/canadian-sovereign-ai-compute-strategy.

Innovation, Science and Economic Development, Canada, Consultation on Copyright in the Age of Generative Artificial Intelligence: What we heard report , 11 February 2025, https://ised-isde.canada.ca/site/strategic-policy-sector/en/marketplace-framework-policy/consultation-copyright-age-generative-artificial-intelligence-what-we-heard-report.

Innovation, Science and Economic Development Canada, The Artificial Intelligence and Data Act (AIDA) – Companion document, February 2023, https://ised-isde.canada.ca/site/innovation-better-canada/en/artificial-intelligence-and-data-act-aida-companion-document.

Innovation, Science and Economic Development Canada, Canadian Sovereign AI Compute Strategy, 15 April 2026, https://ised-isde.canada.ca/site/ised/en/canadian-sovereign-ai-compute-strategy ; Canada, Digital Sovereignty, 31 March 2026, https://www.canada.ca/en/shared-services/services/digital-sovereignty.html.

Innovation, Science and Economic Development Canada, "Canada advances international cooperation on industry, AI and digital technology at the G7 Industry, Digital and Technology Ministers' Meeting ", 9 December 2025, https://www.canada.ca/en/innovation-science-economic-development/news/2025/12/canada-advances-international-cooperation-on-industry-ai-and-digital-technology-at-the-g7-industry-digital-and-technology-ministers-meeting.html.

Innovation, Science and Economic Development Canada, "Minister Solomon, His Majesty King Felipe VI of Spain and Spanish Deputy Prime Minister Cuerpo advance Canada-Spain artificial intelligence cooperation through new memorandum of understanding ", 20 May 2026, https://www.newswire.ca/news-releases/minister-solomon-his-majesty-king-felipe-vi-of-spain-and-spanish-deputy-prime-minister-cuerpo-advance-canada-spain-artificial-intelligence-cooperation-through-new-memorandum-of-understanding-856857718.html .

Innovation, Science and Economic Development Canada , "Canada and Germany sign AI joint declaration and launch Sovereign Technology Alliance ", 14 February 2026, https://www.canada.ca/en/innovation-science-economic-development/news/2026/02/canada-and-germany-sign-ai-joint-declaration-and-launch-sovereign-technology-alliance.html.

Innovation, Science and Economic Development, Canada, "Enabling large-scale sovereign AI data centres ", 16 February 2026, https://ised-isde.canada.ca/site/ised/en/enabling-large-scale-sovereign-ai-data-centres.

Inuit Tapiriit Kanatami, National Inuit Strategy on Research, 2018, https://itk.ca/projects/national-inuit-strategy-on-research.

Irion, Kristina. "The General Data Protection Regulation though the lens of digital sovereignty ", EU Cyber Direct Research Seminar : EU Digital Sovereignty: From Narrative to Policy, 18 March 2022, https://www.ivir.nl/publicaties/download/Irion-GDPR-and-Digital-Sovereignty-

11Mar22.pdf.

Irion, Kristina. "Government Cloud Computing and National Data Sovereignty", (2021) 4:3-4 Policy & Internet 40.

Johnson, Peter & Scassa, Teresa. " Who owns the map? Data sovereignty and government spatial data collection, use, and dissemination ", *Transactions in GIS*, 27, 275–289. https://doi.org/10.1111/tgis.13024.

Karanicolas, Michael. "U.S. social media platforms are a national security threat", *Policy Options*, 8 March 2025, https://policyoptions.irpp.org/2025/03/social-media-threat/.

Kaska, Kadri; Beckvard, Henrik and Minárik, Tomás. "Huawei, 5G and China as a Security Threat", CCDCOE, 2019, https://ccdcoe.org/uploads/2019/03/CCDCOE-Huawei-2019-03-28-FINAL.pdf .

Kim, Dongwoo. "Towards Canada's Asia Strategy on AI", Asia Pacific Foundation of Canada, October 8, 2019 , https://www.asiapacific.ca/publication/towards-canadas-asia-strategy-ai.

KPMG, Trust, attitudes and use of Artificial Intelligence: A global study 2025, https://kpmg.com/xx/en/our-insights/ai-and-technology/trust-attitudes-and-use-of-ai.html.

Law Commission of Ontario and Ontario Human Rights Commission, Human Rights AI Impact Assessment, November 2024, https://www.ohrc.on.ca/en/human-rights-ai-impact-assessment.

Law Society of British Columbia, Innovation Sandbox, https://www.lawsociety.bc.ca/about-us/priorities/innovation-sandbox/.

Law Society of Ontario, Access to Innovation, https://lso.ca/about-lso/access-to-innovation/about-a2i.

Leeson, Howard. "Constitutional Jurisdiction Over Health and Health Care Services in Canada", Discussion Paper No. 12, Government of Canada, August 2022, https://publications.gc.ca/collections/Collection/CP32-79-12-2002E.pdf/.

Lopez Steven, Benjamin. "Feds reveal 6 pillars for long-touted, repeatedly delayed national AI strategy", CBC News, April 28, 2026, https://www.cbc.ca/news/politics/ai-strategy-pillars-evan-solomon-9.7180418 .

Lunau, Kate. "Evan Solomon Wants Canada to Trust AI. Can We Trust Evan Solomon?", The Walrus, January 15, 2026, https://thewalrus.ca/evan-solomon-ai/.

Mann, Hamilton. "AI Homogenization Is Shaping the World", Forbes, 1 November 2024, https://www.forbes.com/sites/hamiltonmann/2024/03/05/the-ai-homogenization-is-shaping-the-world/.

Martin, Alexander. "The AI Threat to Canadian Democracy: Fighting for Digital Sovereignty", Open Canada, 26 May 2025, https://opencanada.org/the-ai-threat-to-canadian-democracy-fighting-for-digital-sovereignty/.

Martin, Alexander. "The threat of US-centric social media", Open Canada, 31 March 2026, https://opencanada.org/the-threat-of-us-centric-social-media/.

McIntyre, Catherine. "New data shows just how dependent Canadian businesses are on U.S. tech", The Logic, 10 March 3036, https://thelogic.co/news/canada-tech-sovereignty-buy-canadian/.

Métis National Council, Métis National Data Strategy , 15 October 2015 , https://www.metisnation.ca/wp-content/uploads/2025/12/Consolidated_Me_tis_Nation_Data_Strategy_Final_15_10_2025_3_56.pdf.

Mehdi, Tahsin and Frenette, Marc. "Exposure to artificial intelligence in Canadian jobs: Experimental estimates", Statistics Canada, 25 September 2024, https://www150.statcan.gc.ca/n1/pub/36-28-0001/2024009/article/00004-eng.htm.

National Centre for Truth and Reconciliation, Truth and Reconciliation Commission of Canada: Calls to Action, https://nctr.ca/about/truth-and-reconciliation-commission-of-canada-calls-to-action/.

National Foreign Trade Council (US), "Canada's Streaming Policy is Unworkable", 22 May 2026, https://www.nftc.org/canadas-streaming-policy-is-unworkable/.

National Institute of Standards and Technology, "Artificial Intelligence Risk Management Framework (AI RMF 1.0)" (January 2023), http://www.nist.gov/itl/ai-risk-management-framework.

Newman, Warren J. "The Promise and Limits of Cooperative Federalism as a Constitutional Principle", (2016) 76 Supreme Court Law Review (2d) 67.

Office of the Privacy Commissioner of Canada, Principles for responsible, trustworthy and privacy-protective generative AI technologies , 7 December 2023, https://www.priv.gc.ca/en/privacy-topics/technology / artificial-intelligence/gd_principles_ai.

Office of the Privacy Commissioner of Canada, Privacy guidance on facial recognition for police agencies , 2 May 2022, https://www.priv.gc.ca/en/privacy-topics/surveillance/police-and-public-safety/gd_fr_202205/.

Office of the Privacy Commissioner of Canada, Joint investigation of Clearview AI, Inc. by the Office of the Privacy Commissioner of Canada, the Commission d'accès à l'information du Québec, the Information and Privacy Commissioner for British Columbia, and the Information Privacy Commissioner of Alberta , 2 February 2021, https://www.priv.gc.ca/en/opc-actions-and-decisions/investigations/investigations-into-businesses/2021/pipeda-2021-001/.

Office of the Privacy Commissioner of Canada, Joint Investigation of OpenAI OpCo , LLC , 6 May 2026, https://www.priv.gc.ca/en/opc-actions-and-decisions/investigations/investigations-into-businesses/2026/pipeda-2026-002/.

O'Leary, Naomi. "'It's surreal': US sanctions lock International Criminal Court judge out of daily life", The Irish Times, 12 December 2025, https://www.irishtimes.com/world/us/2025/12/12/its-surreal-us-sanctions-lock-international-criminal-court-judge-out-of-daily-life/.

Ontario Information and Privacy Commissioner, *AI Scribes: Key Considerations for the Health Sector* , 28 January 2026, https://www.ipc.on.ca/en/resources/ai-scribes-key-considerations-health-sector.

Ontario Information and Privacy Commissioner, Privacy Complaint Report PI21-00001 , 28 February 2024, https://norma.lexum.com/ipc-cipvp/privacy/en/item/521580/index.do.

Ontario Information and Privacy Commissioner , Privacy Complaint Report PX24-00001 , 11 June 2025, https://norma.lexum.com/ipc-cipvp/privacy/en/item/521985/index.do.

Ontario Information and Privacy Commissioner and Ontario Human Rights Commission, Principles for the responsible use of artificial intelligence , 21 January 2026, https://www.ipc.on.ca/en/media/6116/download?attachment.

Ontario, Responsible Use of Artificial Intelligence Directive , 7 January 2022, https://www.ontario.ca/page/responsibl e -use-artificial-intelligence-directive#section-2.

People's Consultation on AI, https://www.peoplesaiconsultation.ca/ .

Prime Minister of Canada, "Prime Minister highlights Canada's leadership in AI at the One Young World Summit in Montréal ", 20 September 2024, https://www.pm.gc.ca/en/news/news-releases/2024/09/20/prime-minister-highlights-canadas-leadership-ai-one-young-world.

Prousky, Jonah. "Canada must tackle its tech talent 'brain drain' problem", Canadian Affairs, 18 July 2024, https://www.canadianaffairs.news/2024/07/18/canada-must-tackle-its-tech-talent-brain-drain-problem/.

Quebec (Attorney General) v. Canada (Attorney General), [2015] S.C.J. No. 14, 2015 SCC 14, [2015] 1 S.C.R. 693 (S.C.C.).

Reference re Employment Insurance Act (Can.), ss. 22 and 23, [2005] S.C.J. No. 57, 2005 SCC 56, [2005] 2 S.C.R. 669.

Report on Engagement with Indigenous Partners and Report on 2022 Engagement with Indigenous Partners https://www.justice.gc.ca/eng/csj-sjc/pa-lprp/wwh4-cqnae4/index.html.

Ruffalo, John. "A Blueprint for Canada's Digital Sovereignty", 18 February 2026, Build Canada, https://www.buildcanada.com/memos/digital-sovereignty.

Saskatchewan, "Guidelines for Government of Saskatchewan Employees in Using Generative Artificial Intelligence ", (n.d.) https://taskroom.saskatchewan.ca/-/media/project/taskroom/documents/guide/artificial-intelligence-guidelines.pdf.

Saskatchewan Information and Privacy Commissioner , *Checklist for Healthcare Organizations Considering the Use of an AI Scribe*, 8 September 2025, https://oipc.sk.ca/assets/checklist-for-healthcare-organizations-considering-the-use-of-an-AI-scribe.pdf.

Satter, Raphael and Alper, Alexandra. "Exclusive: US orders diplomats to fight data sovereignty initiatives", Reuters, 25 February 2026, https://www.reuters.com/sustainability/boards-policy-regulation/us-orders-diplomats-fight-data-sovereignty-initiatives-2026-02-25/.

Scassa, Teresa. "Administrative Law and the Governance of Automated Decision-Making: A Critical Look at Canada's Directive on Automated Decision-Making", (2021) 54:1 University of British Columbia Law Review 251-298.

Scassa, Teresa. "Regulating AI in Canada: A Critical Look at the Proposed Artificial Intelligence and Data Act", (2023) 101 Canadian Bar Review 1-30, https://cbr.cba.org/inde x .php/cbr/article/view/4817.

Scassa, Teresa. "Sovereignty and the Governance of Artificial Intelligence", (2024) 71 UCLA L. Rev. Disc. 214-229.

Sen, Anindya; Clarke, Amanda; Morales, Juan; Penney, Jonathan, and Scassa, Teresa. "AI Innovation, Safety and Regulation in Canada: Perspectives from the Editorial Committee", in Special Issue on Artificial Intelligence Safety and Public Policy in Canada, (2026) 52:S 1 Canadian Public Policy 1 , https://utppublishing.com/doi/10.3138/cpp.2025-066.

Sexsmith, Gin. "AI can help protected threatened languages – but it can also exploit them", Canadian Geographic, 20 April 2026, https://canadiangeographic.ca/articles/ai-can-help-protect-threatened-languages-but-it-can-also-exploit-them/ . See also the work being done by the Abundant Intelligences project of Indigenous AI, https://www.indigenous-ai.net/abundant/.

Sheppard, Eddie and Coletto, David. "What do Canadians think about AI: A Tool for Progress or a Threat to Stability?", Abacus Data, 19 September 2025, https://abacusdata.ca/what-do-canadians-think-about-ai-a-tool-for-progress-or-a-threat-to-stability .

Standards Council of Canada, Canada contributes to global collaboration on responsible AI at the International AI Standards Summit 2025 , 16 Dece mber 2025, https://scc-ccn.ca/resources/ne w s/canada-contributes-global-collaboration-responsible-ai-international-ai-standards.

Standing Committee on Industry and Technology, " Correspondence from the Honourable François-Philippe Champagne, Minister of Innovation, Science and Industry - Amendments to AIDA - 2023-11-28 ", https://www.ourcommons.ca/content/Committee/441/INDU/WebDoc/WD12751351/12751351/MinisterOfInnovationScienceAndIndustry-2023-11-28-Combined-e.pdf.

Swinton, Katherine. "Federalism Under Fire: The Role of the Supreme Court of Canada" (1992) 55 Law and Contemporary Propblems 121.

Tanner, Brooke, and Kerry, Cameron F. " Can small language models revitalize Indigenous languages? ", Brookings, 19 March 2025, https://www.brookings.edu/articles/can-small-language-models-revitalize-indigenous-languages/.

The Dais and Centre for Media, Technology and Democracy (2023) Submission on the proposed Artificial Intelligence and Data Act, https://dais.ca/wp-content/uploads/2023/12/Submission-on-the-Proposed-Artificial-Intelligence-and-Data- Act.pdf.

Toma, Angelo. "Algorithmic Foreign Influence: Rethinking Sovereignty in the Age of AI", Lawfare,

20 August 2025, https://www.lawfaremedia.org/article/algorithmic-foreign-influence--rethinking-sovereignty-in-the-age-of-ai.

Treasury Board Secretariat, Directive on Privacy Practices, https://www.tbs-sct.canada.ca/pol/doc-eng.aspx?id=18309.

Treasury Board Secretariat, Policy on Privacy Protection, https://www.tbs-sct.canada.ca/pol/doc-eng.aspx?id=12510.

Treasury Board Secretariat, 2026 Review of the Privacy Act: Policy Approaches , 2 April 2026, https://www.canada.ca/en/treasury-board-secretariat/services/access-information-privacy/access-information/privacy-act-modernization-policy-approaches/2026-review-privacy-act-policy-approaches.html.

Treasury Board Secretariat, AI Strategy for the Federal Public Service 2025-2027 , 25 February 2025, https://www.canada.ca/en/government/system/digital-government/digital-government-innovations/responsible-use-ai/gc-ai-strategy-overview.html.

Tusikov, Natasha. "Canada must ensure its digital sovereignty in the face of U.S. threats", Policy Options 8 May 2025, https://policyoptions.irpp.org/2025/05/digital-sovereignty/.

Tworek, Heidi and Wanless, Alicia. "How to Confront Canada's Digital Dependence", 1 July 2025, Centre for International Governance Innovation, https://www.cigionline.org/articles/how-to-confront-canadas-digital-dependence/.

UNESCO, Report of the Independent Expert Group on Artificial Intelligence and Culture, 2025, https://www.unesco.org/sites/default/files/medias/fichiers/2025/09/CULTAI_Report%20of%20the%20Independent%20Expert%20Group%20on%20Artificial%20Intelligence%20and%20Culture%20%28final%20online%20version%29%201.pdf.

Universities Canada, "Canada's Universities: The source of Canada's global AI leadership", February 2026, https://univcan.ca/wp-content/uploads/2026/02/February-2026-The-source-of-Canadas-global-AI-leadership-Policy-primer.pdf.

UN Trade and Development, "AI's $4.8 trillion future: UN Trade and Development alerts on divides, urges action", 7 April 2025, https://unctad.org/news/ais-48-trillion-future-un-trade-and-development-alerts-divides-urges-action.

Usardi, Annalisa and Drut, Bastien. "The diverse economic impacts of artificial intelligence", SUERF Policy Brief No. 1174, 28 May 2025, https://www.suerf.org/publications/suerf-policy-notes-and-briefs/the-diverse-economic-impacts-of-artificial-intelligence/

US Chamber of Commerce, "U.S. Chamber Strongly Condemns Canada's Online Streaming Act", 23 May 2026, https://www.uschamber.com/international/u-s-chamber-strongly-condemns-canadas-online-streaming-act.

von Tigerstrom, Barbara. *Information and Privacy Law in Canada* , (Irwin Law, 2020).

Walter, Maggie; Kukutai, Tahu; Russo Carroll, Stephanie and Rodriguez-Lonebear, Desi, eds., *Indige-*

nous Data Sovereignty and Policy, Taylor Francis, 2021.

Westendarp, Louis. "Lost in translation: Europe's quest to speak AI ", Politico, 19 July 2022, https://www.politico.eu/article/translation-europe-quest-speak-ai-technology/.

White House, Executive Order 14203, Imposing Sanctions on the International Criminal Court, 6 February 2025, https://www.whitehouse.gov/presidential-actions/2025/02/imposing-sanctions-on-the-international-criminal-court/.

White House, Executive Order 14179 – Removing Barriers to American Leadership in Artificial Intelligence , January 23, 2025, https://www.whitehouse.gov/presidential-actions/2025/01/removing-barriers-to-american-leadership-in-artificial-intelligence/.

White House, "Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence", US Executive Order 14110 of October 30, 2023, https://www.federalregister.gov/documents/2023/11/01/2023-24283/safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence.

White House, "Defending American Companies and Innovators From Overseas Extortion and Unfair Fines and Penalties ", 21 February 2025, https://www.whitehouse.gov/presidential-actions/2025/02/defending-american-companies-and-innovators-from-overseas-extortion-and-unfair-fines-and-penalties/.

Yousif, Nadine. "Seven lawsuits filed against OpenAI by families of Canada mass-shooting victims", BBC online, 29 April 2026, https://www.bbc.com/news/articles/c99l03k0ly4o.

Zewe, Adam. "Explained: Generative AI's environmental impact", MIT News, 17 January 2025, https://news.mit.edu/2025/explained-generative-ai-environmental-impact-0117.

Zhang, Lawrence. "The right way for Canada to secure cloud sovereignty", Policy Options, 29 October 2025, https://policyoptions.irpp.org/2025/10/cloud-sovereignty/.

Zhang, Lawrence and Castro, Daniel. "Canada should harness its AI advantage, not squander it", Macdonald Laurier Institute, May 19, 2026, https://macdonaldlaurier.ca/canada-should-harness-its-ai-advantage-not-squander-it-lawrence-zhang-and-daniel-castro/.

Laws, Bills and Treaties

Alberta, *Election Finances and Contributions Disclosure Act* , RSA 2000, c E-2, s . 44.94 99995.

Alberta, *Protection of Privacy Act* , SA 2024, c P-28.5, s. 5 .1 , s. 6.

Bill C-27, An Act to enact the Consumer Privacy Protection Act, the Personal Information and Data Protection Tribunal Act and the Artificial Intelligence and Data Act and to make consequential and related amendments to other Acts, 1st Sess , 44th Parl , 2022 (first reading 16 June 2022), online: <www.parl.ca/DocumentViewer/en/44-1/bill/C-27/first-reading> [perma.cc/]

Canada, *Budget Implementation Act* , 2025, No. 1 , SC 2025, c. 3.

Canada, The Constitution Act, 1867 , 30 & 31 Vict, c 3 .

Canada, *Personal Information Protection and Electronic Documents Act*, S.C. 2000, c. 5.

Canada, *Privacy Act*, RSC 1985, c P-21 .

Canada, *Red Tape Reduction Act*, SC 2015, c 12.

Canada, United Nations Declaration on the Rights of Indigenous Peoples Act , SC 2021, c. 14.

Council of Europe Framework Convention on Artificial Intelligence , 133rd Session of the Committee of Ministers (Strasbourg, 17 May 2024).

Manitoba, Bill 49, The Business Practices Amendment Act, 3rd Sess., 43rd Legislature , 2026, https://web2.gov.mb.ca/bills/43-3/b049e.php .

Manitoba, Bill 49, The Business Practices Amendment Act , 3 rd Sess , 43rd Legislature, https://web2.gov.mb.ca/bills/43-3/b049e.php.

Manitoba, Bill 51, The Public Sector Artificial Intelligence and Cybersecurity Governance Act, 3 rd Sess. 43rd Legislature, https://web2.gov.mb.ca/bills/43-3/b051e.php.

Nova Scotia, *Personal Information International Disclosure Protection Act*, https://www.canlii.org/en/ns/laws/stat/sns-2006-c-3/96537/sns-2006-c-3.html.

Ontario, Working for Workers Four Act , 2024, SO 2024, c 3 , s. 8.4.

Quebec Civil Code, CQLR c CCQ-1991.

Ontario, Working for Workers Four Act , 2024, SO 2024, c 3 , s. 8.4.

Proposal for a Regulation of The European Parliament and of the Council Laying Down Harmonised Rules On Artificial Intelligence (Artificial Intelligence Act) And Amending Certain Union Legislative Acts, (2021), EUR-Lex < eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52021PC0206 > [perma.cc/] .

EU Artificial Intelligence Act, Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act).

Quebec, *Act respecting Access to documents held by public bodies and the Protection of personal information* , CQLR c A-2.1, s. 65.2.

Quebec, An Act to modernize legislative provisions as regards the protection of personal information, SQ 2021, c 25 , s. 27, s. 111.

UN General Assembly, United Nations Declaration on the Rights of Indigenous Peoples : resolution / adopted by the General Assembly, A/RES/61/295, 2 October 2007, https://www.refworld.org/legal/resolution/unga/2007/49353.

U.S., CLOUD Act, H.R.4943 — 115th Congress (2017-2018).

AI REGULATION IN LATIN AMERICA

Rodrigo Soares Peva
Mestre em Direito - UERJ
Servidor Público Federal – TRF2
DOI: https://doi.org/10.59224/dti8.ch8

Abstract: This article examines the normative instruments regulating artificial intelligence in ten Latin American countries, based on a systematic documentary survey conducted through August 2026. Under a descriptive-comparative approach, it organises them by generality, institutional origin, and legal nature. The mapping reveals the predominance of strategic plans, guidelines, and administrative acts, while comprehensive statutory frameworks remain under legislative consideration. The region is undergoing a gradual institutionalisation, marked by asymmetries and differing speeds of normative maturation.

Keywords: AI regulation; Latin America; Digital sovereignty

Table of contents: 1. Introduction 2. Overview of AI Regulation 3. Criteria for Categorizing Latin American Regulatory Instruments 4. Regulatory Mapping in Latin America 4.1 Argentina 4.2 Brazil 4.2.1 Brazil's National AI Plan – PBIA 4.2.2 Brazil's multilevel AI governance 4.2.3 Subnational lawmaking 4.2.4 Sectoral governance: judiciary, elections, and criminal law 4.2.5 A broadened conception of AI sovereignty 4.3 Chile 4.4 Colombia 4.5 Costa Rica 4.6 Ecuador 4.7 Mexico 4.8 Panama 4.9 Peru 4.10 Uruguay 5. 2023 Santiago Declaration 6. What Can Be Learned from the Latin American Regulatory Trend? Conclusion. References

1. INTRODUCTION

This study examines the normative instruments regulating artificial intelligence

in ten Latin American countries: Argentina, Brazil, Chile, Colombia, Costa Rica, Ecuador, Mexico, Panama, Peru, and Uruguay. Its aim is to identify the normative patterns, institutional arrangements, and governance strategies that characterise the region. The research adopts a descriptive-comparative approach, based on a systematic documentary survey that draws primarily on official normative sources and on an international database monitoring AI regulation, complemented, where necessary, by secondary literature and press reports for the most recent developments.

Rather than describing each instrument in isolation, the study adopts a classificatory perspective, on the premise that any assessment of the regional landscape must be preceded by a typological organisation of the existing acts. To that end, normative and policy documents are arranged according to three structuring criteria: general versus sectoral regulation; rulemaking by the Executive versus legislative initiative; and the legal nature of the instrument, whether statute, decree, resolution, guideline, strategic plan, or administrative guide. This classification serves to locate where regulatory choices are effectively concentrated, whether at the strategic level, in administrative regulation, or in formal legislation.

Only after this systematisation does the study turn to an objective presentation of each country, setting out the main instruments in each jurisdiction, their stage of implementation, and their declared priorities. The approach avoids fragmented description and enables an integrated reading of the Latin American experience while preserving the institutional specificities of each legal system. The comparison is consolidated in two tables: the first summarising, for each country, the instruments in force or under discussion, their legal nature, the issuing branch, the regulatory approach, the existence of a dedicated AI authority, and a distinctive feature; the second grouping the guiding principles that recur across the region.

Two limitations should be noted. Given the pace at which the field evolves, the analysis reflects the state of regulation as of 1 August 2026 and does not purport to be exhaustive. Further instruments may be under discussion or may have emerged after the cut-off date. The study is also confined to instruments formally published or registered in the sources consulted, and does not encompass informal practices, undocumented drafts, or private initiatives unconnected to normative acts.

Beyond systematisation, the study engages with the scholarship on AI govern-

ance, systemic regulation, sovereignty, and institutional design. The regulatory documents are read in light of these contributions, not to impose external models or to anticipate conclusive judgments, but to assess how far the Latin American experiences reveal a consistent and institutionally sustainable regulatory trajectory. The comparative examination ultimately asks whether it is already possible to discern structuring guidelines capable of shaping a relatively stable governance path, or whether the present stage remains one of normative experimentation and institutional fragmentation, counselling greater analytical caution.

2. OVERVIEW OF AI REGULATION

Governance refers to the idea of a plurality of actors who do not possess the power or authority to decide or implement solutions or policies unilaterally and directly, as noted by William Walters. [1] In this sense, governance concerns the interactions between public and private actors, such as manufacturers, engineers, lawyers, users, civil society representatives, and researchers, with the aim of addressing diverse social problems.

The regulation of artificial intelligence may also be understood as a manifestation of digital sovereignty.[2] Teresa Scassa observes that, in the technological environment, sovereignty extends beyond its traditional territorial dimension and comes to express the State's capacity to exercise control and regulatory autonomy over the digital sphere. In this context, AI governance constitutes both an internal normative act and an assertion of power at the international level, against a backdrop of competition over the definition of the norms and standards that will govern the technology. [3] This perspective is particularly relevant in Latin America, as it makes it possible to examine the extent to which national regulatory models reflect their own particularities and normative choices or, conversely, reproduce or adapt foreign regulatory

---

1 WALTERS, William. Some critical notes on "governance." Studies in Political Economy, v. 73, 2004. Disponível em: https://www.tandfonline.com/doi/abs/10.1080/19187033.2004.11675150. Acesso em: 04 mar. 2025, *passim*.

2 SCASSA, Teresa. Sovereignty and the governance of artificial intelligence. *UCLA Law Review*, Los Angeles, v. 71, n. 1, p. 214-229, maio 2024, p. 216-217.

3 *Ibid*.

models, especially in light of the extraterritorial effects of European regulation.[4]

Judge, Nitzberg and Russell distinguish governance from regulation, defining the former in terms of collective decision-making involving multiple actors and associating its development with the proliferation of different policy instruments and actor groups. [5]This distinction is particularly relevant to the Latin American context, where AI governance may emerge through legislation, national strategies, institutional frameworks, guidelines, or other policy instruments.[6]

It is essential not to lose sight of the fact that governance mechanisms should be inclusive, transparent, multidisciplinary, multilateral (including the possibility of mitigation and redress beyond national borders), and multi-stakeholder. [7] Through collaboration among different stakeholders, including marginalized groups from both urban and rural areas, international organizations, transnational corporations, and academic institutions, it becomes possible to develop laws, regulations, and public policies that can be tested in safe environments before being formally adopted by the State. [8]

Such pluralistic cooperation is justified insofar as regulators still lack sufficient expertise to effectively oversee the development of AI systems. [9] It is therefore crucial that developers and programmers act transparently and collaborate with experts from multiple fields to share knowledge and ideas. This exchange of knowledge can

---

4 *Ibid*.

5 JUDGE, Brian; NITZBERG, Mark; RUSSEL, Stuart. When code isn't law: rethinking regulation for artificial intelligence. *Policy and Society*. Oxford: Oxford University Press, p. 01-13, maio 2024, p. 2.

6 *Ibid*.

7 UNESCO. Recomendação sobre a Ética da Inteligência Artificial. Paris: UNESCO, 2022. Disponível em: https://unesdoc.unesco.org/ark:/48223/pf0000381137_por. Acesso em: 02 de mar. de 2025, p. 27.

8 UNESCO. Recomendação sobre a Ética da Inteligência Artificial. Paris: UNESCO, 2022. Disponível em: https://unesdoc.unesco.org/ark:/48223/pf0000381137_por. Acesso em: 02 de mar. de 2025, p. 27.

9 GUIHOT, Michael; MATTHEW, Anne F.; SUZOR, Nicolas. Nudging Robots: Innovative Solutions to Regulate Artificial Intelligence. Vanderbilt Journal of Entertainment & Technology Law, no prelo, 28 jul. 2017. Disponível em: https://ssrn.com/abstract=3017004. Acesso em: 02 mar. 2025, *passim*.

contribute to a more informed regulatory framework that benefits society as a whole.[10]

One feature identified in some Latin American approaches to AI governance is the use of regulatory sandboxes. Harasimiuk and Braun describe this mechanism as an experimental and adaptive regulatory approach, in which new, relatively non-intrusive rules are tested alongside emerging technologies before being introduced on a broader and binding basis.[11] Rather than allowing technological development to proceed without regulatory oversight, this model enables regulation to accompany innovation from the testing stage.[12] The authors further emphasize that regulatory sandboxes should involve multiple public and private stakeholders, allowing new technologies to be assessed without unnecessarily hindering innovation, while simultaneously safeguarding public and private interests and preventing unacceptable risks.[13]

The very understanding of the actors responsible for artificial intelligence governance may vary depending on the regulatory approach adopted. In this regard, Frank Wolf draws attention to the greater presence of regulatory bodies in the public debate on AI governance and ethics, in contrast to the relatively limited participation of private actors who develop, invest in, and operate these technologies, while emphasizing that both sectors pursue good governance.[14] This perspective provides a useful lens through which to understand the approach currently advocated by the Argentine government, which, under President Javier Milei, has expressed a preference for an environment characterized by limited regulatory intervention in artificial intelli-

---

[10] GUIHOT, Michael; MATTHEW, Anne F.; SUZOR, Nicolas. Nudging Robots: Innovative Solutions to Regulate Artificial Intelligence. Vanderbilt Journal of Entertainment & Technology Law, no prelo, 28 jul. 2017. Disponível em: https://ssrn.com/abstract=3017004. Acesso em: 02 mar. 2025, *passim*.

[11] HARASIMIUK, Dominika Ewa; BRAUN, Tomasz. Regulating artificial intelligence: binary ethics and the law. Abingdon: Routledge, 2021, p. 138-139.

[12] *Ibid*.

[13] *Ibid.*

[14] Frank wolf:WOLF, Frank. *A blueprint for the regulation of artificial intelligence technologies*. Bradford: Ethics International Press, 2022, p. 14.

gence.[15] While Wolf's argument should not be construed as an endorsement of deregulation, the Argentine experience illustrates a potential redistribution of the leading role in AI governance, with greater prominence being afforded to private actors and technological development in relation to state regulatory intervention.[16]

For Lucas Introna[17], governmentality provides a lens through which to observe the performative outcomes of governing practices. He further notes that Foucault prefers the notion of governmentality, or the "*conduct of conduct*", because it extracts the "how" of governance rather than merely the "what." [18] In other words, governmentality is concerned with examining how practices and technologies within governance structures influence individuals and groups in their actions.

Eduardo Magrani[19], in turn, proposes new ontological and epistemological perspectives on artificial intelligence, reflecting on the phenomenon of "*distributed irresponsibility*," potentially produced by current legal frameworks that have not yet fully adapted to technological evolution. This gap compromises the establishment of causal links between agents (human and/or non-human) and harms within complex sociotechnical systems in which human beings and technical components interact in intricate ways.

Ashish Saxena's study conducted a literature review of thirty articles addressing AI governance across various sectors in 2024 and found that excessive reliance on AI may weaken human judgment, thereby requiring robust governance structures. The research reached two main conclusions: (a) predictive analytics enables proactive policymaking, although algorithmic bias and privacy concerns persist; and (b) there is a need to integrate public feedback into AI governance processes. However, the

---

15 *Ibid*.

16 *Ibid*.

17 INTRONA, Lucas D. Algorithms, governance, and governmentality: on governing academic writing. Science, Technology, & Human Values, v. 41, n. 1, p. 17–49, 2016. Disponível em: http://www.jstor.org/stable/43671281. Acesso em 02 mar. 2025, *passim*

18 INTRONA, Lucas D. Algorithms, governance, and governmentality: on governing academic writing. Science, Technology, & Human Values, v. 41, n. 1, p. 17–49, 2016. Disponível em: http://www.jstor.org/stable/43671281. Acesso em 02 mar. 2025,*passim*

19 MAGRANI, Eduardo. Novas perspectivas sobre ética e as leis da inteligência artificial. Internet Policy Review, v. 8, n. 3, p. 1-19, set. 2019. DOI: 10.14763/2019.3.1420, *passim.*

author acknowledges the study's limitations, particularly its inability to assess the long-term impact of AI on governance. [20]

Baldwin, Cave, and Lodge identify three central challenges in measuring regulatory quality. The first concerns the definition of the relevant criteria and objectives and the weight assigned to each of them, since different actors may adopt distinct conceptions of regulatory quality.[21] The second relates to delimiting the object of assessment itself, which may focus on the resources employed, the processes adopted, the outcomes achieved, or the impacts actually produced by the regulation.[22] The third refers to identifying whose performance is to be assessed, particularly in light of decentralized, polycentric, or network-based regulatory models, in which different actors, instruments, and policies jointly contribute to the production of regulatory outcomes.[23]

Although not specifically focused on Latin America, Kartskhiya's observation that "*there are no unified approaches to regulating artificial intelligence and robotics technologies worldwide*" is also reflected in the ten jurisdictions examined in this study.[24] The regulatory landscape across these countries is marked by significant diversity, as each jurisdiction adopts distinct legal, policy, or institutional approaches to artificial intelligence, even though not all of them have enacted comprehensive AI-specific legislation.[25] This heterogeneity reinforces the author's broader point that the regulation of AI remains fragmented and shaped by unresolved legal, technological, and institutional questions.[26]

---

[20] SAXENA, Ashish K. AI in Governance and Policy Making. International Journal of Science and Research (IJSR), v. 13, n. 5, p. 1218-1223, maio 2024. Disponível em: https://www.ijsr.net/getabstract.php?paperid=SR24519015426. DOI: https://www.doi.org/10.21275/SR24519015426. Acesso em: 02 de mar. de 2025., *passim*.

[21] BALDWIN, Robert; CAVE, Martin; LODGE, Martin. *Understanding Regulation: Theory, Strategy, and Practice*. Oxford: Oxford University Press, 2012, p. 34-36.

[22] *Ibid.*

[23] *Ibid*.

[24] KARTSKHIYA, Aleksandr Amiranovich. Legal horizons of the new artificial intelligence paradigm. *Legal Issues in the Digital Age*, Moscow, v. 5, n. 2, p. 4-35, jul. 2024.

[25] *Ibid.*

[26] *Ibid*.

3 CRITERIA FOR CATEGORIZING LATIN AMERICAN REGULATORY INSTRUMENTS

This section sets out the criteria used to categorize the regulatory instruments adopted across ten Latin American countries in the field of artificial intelligence. Rather than an exhaustive account of each national system, it seeks to map the main instruments, whether enacted, pending, or still at the stage of policy, and to identify convergences and divergences across the region. The selection privileges instruments specifically directed at AI, while treating adjacent frameworks, such as data-protection statutes, as supporting infrastructure rather than as AI regulation in the strict sense. Given the rapidly evolving nature of the field, the analysis reflects the state of regulation as of 1 August 2026 and draws primarily on official normative sources, complemented, where necessary, by secondary literature and, occasionally, by press sources for the most recent developments.

The comparison is structured around the dimensions summarised in Table[27] , chosen to be simple enough for a text of limited length yet sufficient to capture the relevant institutional choices in each country: (i) the normative instruments in force or under discussion; (ii) their legal nature, ranging from enacted statutes to bills, strategies, and administrative acts; (iii) the issuing branch, Executive, Legislative, or Judiciary. reflecting where regulatory initiative has concentrated; (iv) the regulatory approach, notably the adoption of risk-based or principle-based models; (v) the existence of a dedicated AI authority; and (vi) each country's distinctive feature. On this basis, the analysis describes the predominant elements in each jurisdiction, with attention to the stage of regulatory consolidation.

The instruments compiled in Table correspond to the most recent normative and policy documents identified for each country up to the research cut-off date. Given the speed at which this field evolves and the multiplicity of initiatives underway across the region, often at different institutional levels and stages of discussion, the table does not purport to be exhaustive. Other instruments may be under debate or

[27] The table's layout was produced with the assistance of an AI tool; the data were collected and entered manually by the author, who is responsible for their accuracy.

may have emerged after the cut-off date without being captured here. The aim is therefore to offer a representative rather than a complete picture, sufficient to identify the region's main regulatory trends.

The instruments surveyed converge around a recurring core of guiding principles , transparency, non-discrimination and equity, human oversight, risk prevention and proportionality, among others, often expressed through different wording across jurisdictions. This convergence should not, however, be read as an argument for a purely principle-based model of regulation. Principles, on their own, are open-textured and insufficient to resolve concrete cases; their effectiveness depends on subsequent concretization through binding rules, oversight authorities, and enforcement mechanisms. Their shared presence is nonetheless significant. It reveals a common normative grammar across the region and offers a starting point for practical implementation, against which more detailed regulatory choices can be developed and assessed.

Overview of AI Regulation in Latin America — Comparative View by Country						
Country	**Normative instruments**	**Legal nature**	**Issuing branch**	**Regulatory approach**	**AI authority**	**Distinctive feature**
Peru	AI Law No. 31,814 (2023) + Regulation (Supreme Decree 115-2025-PCM); Law No. 1412 of 2018 on Digital Government; National Artificial Intelligence Strategy 2026–2030; National Digital Transformation Policy to 2030 (2023) and Law 32,314/2025 (Criminal Code reform on AI)	Law in force + Supreme Decree	Legislative + Executive	Risk-based (2 tiers):Misuse and High-risk use	SGTD	Regional pioneer with a general AI law already regulated; AI as a criminal aggravating factor (Law 32,314) and Govt. recognises rules alone don't ensure coherent state AI adoption (Strategy 2026-2030)
Chile	Cybersecurity Framework Law 21,663 (2024); New Data Law 21,719 (ieffective as of December 1, 2026); AI Bill (Bulletin 16,821-19); 2024 National AI Policy	Related laws in force + AI bill	Legislative	Risk-based (4 tiers, in the bill): Unacceptable risk, high risk, evident risk and limited risk	National Cybersecurity Agency (existing); Technical Advisory Council on AI (in the bill)	Neurorights in the Constitution; mature regulatory infrastructure (cyber + data protection) and Right to explainability and human intervention (New Data Protection Law)
Brazil	AI Bill 2,338/23 (pending); CNJ Resolution 615/2025 (replacing Resolution. 332/2020); PBIA (2024-2030); TSE Resolutions 23,610/2019 and 23.755/2026; Federal Law 15,123/25 (Penal Law), 15,211/2025 (Digital ECA) and 14,129/2021 (Digital Government) and Law 13,709/2018 (Data Protection); Municipality of Rio de Janeiro, AI Law 8,926/2025; Goiás, AI Compl. Law 205/2025; Paraná, AI Law 22,343/2025 and Alagoas, AI Law 9,095/2023	Bill + Judiciary norm + public policy + State and municipal AI laws + Supporting legal framework	Legislative / Judiciary / Executive	Risk-based (2 tieres, in the bill): excessive risk and high risk	the National Data Protection Authority (ANPD)	Strong sector-specific governance in the Judiciary (CNJ); emphasis on technological sovereignty; state-level laws
Colombia	AI Bill 422/2025; National AI Policy/CONPES 4144 (2025); Judiciary Guide (2024), Law 1,581/2012 (Data Protection); National Digital Strategy of Colombia (2023-2026), Roadmap for the Development and Application of Artificial Intelligence (2024); Superior Council of the Judiciary: AI Guidelines (2024) and	Bill + public policy + Judiciary guide	Legislative / Executive / Judiciary	Risk-based (4 tiers, in the draft): prohibited, high risk, limited risk and low risk.	Ministry of Science, Technology and Innovation (in the AI bill)	First country to implement the UNESCO recommendation in the Judiciary; distributed governance
Mexico	Pending amendment to Art. 73, XVII of the Constitution, to include AI in Congress's legislative powers; 2018 National Artificial Intelligence Strategy. General AI Bill (2026); sectoral reforms (Copyright, Labor Law, Criminal Code and 2025 Protection Data);	Bill + sectoral reforms in force	Legislative	Risk-based (2 tiers: proibited aplications and high risk) + sectoral	National AI Authority (in the AI bill)	Sector-by-sector regulation; technological sovereignty; protection of artists' voice/image against AI and neural rights (In the comprehensive AI bill- 223 articles).
Costa Rica	National AI Strategy (2024-2027); Bill 23,771/2023	Public policy + bill	Executive / Legislative	Risk-based (4 tiers, inspired by the EU AI Act: unacceptable, high, medium, and low risk)	National Agency for Digital Government (NAD G) in the National Strategy	Strategy explicitly modeled on the EU AI Act; bill drafted with AI assistance; R&D&I and focus on specific regulations, decrees, and guidelines (National Strategy)
Ecuador	EFIA-EC Strategy (2025); Regulation bill (2024); Ethical Decalogue (2025)	Public policy + bill + ethics code	Executive / Legislative	Risk-based (4 tiers, in the bills): Low, Moderate risk , High risk and Extreme risk	National AI Oversight Authority (in the AI bill)	Sectoral ethical decalogue
Uruguay	Tratado del Consejo de Europa sobre IA; Montevideo Declaration; (2024); National AI Strategy (2024-2030); legal basis in Law 20,212, art. 74.	Public policy (with legal basis)	Executive (AGESIC) / Legislative	Principle-based and Internation commitments	AGESIC	First Latin American country to accede to the Council of Europe Convention on AI (2025)
Argentina	Law 25,326/2000; 2019 National AI Plan; provincial court protocols; Buenos Aires Bar Association guideline and 2026 General Companies Bill	Public policy + court norms + soft law	Executive / Provincial Judiciary / entities	Principle-based; no active law or bill	No specific authority	Executive's deregulatory stance; no active bill; provincial judicial protocols and automated companies (in General Companies Bill)
Panama	Bill 588 (approved in first debate, 2026); PENCYT 2025-2029; AI 2036 (National Artificial Inteligence Strategy); National Agenda for Advanced Technologies.	Bill + public policy	Legislative / Executive	Adaptive governance (risk + experimentation) high risk and low risk (AI 2036 - National Artificial Inteligence)	National Commission on Critical and Emerging Technologies (AI Subcommittee)	Focus on a digital hub and AI as an economic driver; adaptive governance and agile and experimentation principle.

4 REGULATORY MAPPING IN LATIN AMERICA

4.1 ARGENTINA

Within its national science, technology and innovation policy, "*Argentina Innovadora 2030"* (AI2030) places technological capabilities as part of a wider project of productive transformation.[28] Its objectives connect the strengthening of science and technology with a more knowledge-intensive economy, while also addressing territorial development, employment, sustainability, and Argentina's position in the international economy.[29] The OECD Principles on Artificial Intelligence serve as an important reference for this framework, although the strategy envisages their application in a manner responsive to the country's own circumstances.[30]

This policy background coexists with a long-standing data protection framework. Law No. 25,326 of 2000[31] follows international data protection standards and has been recognised by the European Commission as providing an "*adequate level of protection*".[32] More recently, however, the political discourse surrounding AI regulation has emphasised regulatory restraint. President Milei has referred to the importance of avoiding the "*deadly hand of premature and misguided regulation*," reflecting a markedly cautious stance towards early regulatory intervention in the development of AI.[33]

---

[28] ARGENTINA. Presidencia de la Nación. *Plan Nacional de Inteligencia Artificial*. Buenos Aires: Presidencia de la Nación, 2019. Disponível em: https://api.oecdai.org/storage//policy-initiatives/Jul2025/fu_rh7qgrkz9xkx89g.pdf. Acesso em: 6 ago. 2026.

[29] *Ibid*.

[30] *Ibid*.

[31] ARGENTINA. *Ley nº 25.326, de 4 de octubre de 2000*. Protección de los Datos Personales. Buenos Aires: Congreso de la Nación Argentina, 2000. Disponível em: https://servicios.infoleg.gob.ar/infolegInternet/anexos/60000-64999/64790/texact.htm. Acesso em: 6 ago 2026;

[32] UNIÃO EUROPEIA. Comissão Europeia. *Adequacy decisions*. Brussels: European Commission, [s. d.]. Disponível em: https://commission.europa.eu/law/law-topic/data-protection/international-dimension-data-protection/adequacy-decisions_en. Acesso em: 6 ago. 2026.

[33] MILEI defende IA desregulada após alerta de historiador Harari. *UOL Notícias*, [São Paulo], 8 jun. 2026. Disponível em: https://noticias.uol.com.br/ultimas-noticias/afp/2026/06/08/milei-defende-ia-desregulada-apos-alerta-de-historiador-harari.htm. Acesso em: 6 ago. 2026;

In parallel, more concrete governance initiatives have emerged within the judiciary, particularly through soft-law instruments adopted at the provincial level. Between October 2024 and April 2025, courts in Río Negro, San Juan, Santa Fe, and Jujuy adopted protocols governing the use of generative AI.[34] These instruments apply, according to the respective initiatives, to judges, officials, and court employees, illustrating the development of institution-specific rules for the use of AI even in the absence of a single overarching regulatory instrument.[35]

Professional bodies have also begun to develop their own standards for the use of AI. In July 2025, the Buenos Aires Bar Association adopted two instruments addressing its implications for legal practice. [36]A guide for lawyers and criteria for its ethical and responsible use. Rather than treating AI as a substitute for professional expertise, these documents emphasise the continued role of human judgment and legal reasoning, together with the need for human supervision.[37] They also address the handling of sensitive case information and the importance of keeping lawyers adequately trained and up to date.[38] The guidance extends to practical aspects of interacting with AI systems, including prompt formulation.[39]

Corporate law has likewise become a field for emerging AI-related legislative proposals. The proposed reform of the 1972 General Companies Law provides, in Article 4, for the category of an "*Automated Company*", encompassing companies whose ordinary activities are carried out through "*autonomous algorithmic systems or AI agents without reliance on employees or other human resources*". [40]The proposal

---

[34] OXFORD INSTITUTE OF TECHNOLOGY AND JUSTICE. *Argentina*. Oxford: University of Oxford, 2026. Disponível em: https://www.techandjustice.bsg.ox.ac.uk/research/argentina. Acesso em: 6 ago. 2026

[35] *Ibid*.

[36] *Ibid*.

[37] *Ibid*.

[38] *Ibid*.

[39] *Ibid*.

[40] ARGENTINA. Poder Ejecutivo Nacional. *Mensaje: Ley General de Sociedades*. Buenos Aires, 2026. Projeto de lei encaminhado ao Honorable Congreso de la Nación. Disponível em: https://aldiaargentina.microjuris.com/wp-content/uploads/2026/06/proyecto-ley-general-sociedades.pdf?ref=digital-nomos.ghost.io. Colegio Público de la Abogacía de la Capital Federal – documento em PDF. Acesso em: 6 ago. 2026.

also contemplates in Article 102 the "*use of AI and algorithmic systems*" by corporate management bodies for operational tasks and decision-making. [41]Such use, however, does "*not relieve administrators of responsibility or of their obligations to configure and oversee the systems and the results they produce*".[42]

4.2 BRAZIL

4.2.1. Brazil's National AI Plan – PBIA

Beyond its legislative framework, Brazil articulates its strategic vision through the Brazilian Artificial Intelligence Plan (PBIA), tellingly subtitled "*AI for the good of all*" (*IA para o bem de todos*).[43] The Plan adopts a human-centered approach, aligned with national interests and the right to development, and oriented toward addressing the country's social, environmental, and economic challenges. [44]This vision rests on five pillars[45]operationalized through five axes, spanning infrastructure, education and training, public services, business innovation, and support for AI regulation and governance.[46]

---

41 *Ibid*.

42 *Ibid*.

43 BRASIL. Ministério da Ciência, Tecnologia e Inovação. *Plano Brasileiro de Inteligência Artificial (PBIA): IA para o bem de todos*. Brasília, DF: MCTI, 2024. Disponível em: https://www.gov.br/mcti/pt-br/acompanhe-o-mcti/transformacaodigital/plano-brasileiro-de-inteligencia-artificial-pbia-_vf.pdf. Acesso em 02 agoo. 2026.

44 *Ibid*.

45 Five pilars: "i) human-centered and accessible to all, grounded in dignity, social rights, and cultural diversity, while preventing inequality and discriminatory bias; (ii) oriented toward overcoming social, environmental, and economic challenges, contributing to the Sustainable Development Goals; (iii) founded on the right to development and national sovereignty, fostering technological autonomy and economic competitiveness; (iv) transparent, traceable, and accountable, safeguarding privacy, data sovereignty, cybersecurity, consumer protection, and intellectual property; and (v) globally cooperative on fair and mutually beneficial terms, advancing the integrity of information and the defense of democracy."

46 BRASIL. Ministério da Ciência, Tecnologia e Inovação. *Plano Brasileiro de Inteligência Artificial (PBIA): IA para o bem de todos.* Brasília, DF: MCTI, 2024. Disponível em: https://www.gov.br/mcti/pt-br/acompanhe-o-mcti/transformacaodigital/plano-brasileiro-de-

Axis 5 ("*Support for the Regulatory Process and AI Governance*") is backed by the Brazilian Observatory on Artificial Intelligence (OBIA) and the National Center for Algorithmic Transparency and Trustworthy AI. [47]Through the coordinated action of these two initiatives, it will be possible to develop guidelines for the ethical use of artificial intelligence and to disseminate advances produced by a network of Brazilian and international researchers, scientists, legal scholars, engineers, and programmers in seminars, conferences, and forums supported by the country.[48]

To achieve the goals established under these axes, substantial investment will be required in Brazil, similar to the efforts undertaken by the United States between 2021 and 2024, when approximately BRL 63 billion in public resources were invested in research and development (R&D).[49] It is estimated that the European Union, in turn, will invest around BRL 16 billion between 2024 and 2027 to establish "*AI Factories*" and foster the application of AI in industrial and social sectors.[50]

Brazil is expected to invest BRL 23.03 billion between 2024 and 2028, including BRL 435.04 million in immediate-impact actions and BRL 103.25 million dedicated specifically to supporting the regulatory process and AI governance.[51] It is worth noting that funding sources vary, ranging from BRL 12.72 billion in credit lines (FNDCT/FINEP + BNDES + others) to BRL 1.06 billion originating from the private sector (through investments and counterpart funding).[52]

One of the PBIA's more meaningful contributions is its acknowledgment that AI governance need not await a future, purpose-built "*AI law*." The Plan emphasizes that, while Bill 2,338/2023 awaits approval, the country can already draw on a relevant normative infrastructure for technological governance, precisely because AI systems depend so heavily on data. That base comprises the Access to Information Law, which sustains the Open Data Policy; the General Data Protection Law, which sets parameters for personal data processing and protects fundamental rights; and

inteligencia-artificial-pbia-_vf.pdf. Acesso em 02 ago. 2026.

47 *Ibid*.

48 *Ibid.*

49 *Ibid.*

50 *Ibid.*

51 *Ibid.*

52 *Ibid.*

Decree 10,046/2019, which regulates shared data use across the federal public administration. [53]On this reading, governance is anchored in existing rules on data access, transparency, protection, and sharing, rather than in a single instrument still to be enacted.[54]

In the Brazilian case, AI governance is still primarily structured through planning instruments and administrative coordination mechanisms. The Brazilian Artificial Intelligence Strategy (EBIA) and the Brazilian Artificial Intelligence Plan 2024–2028 function as guiding frameworks, implemented through decrees, resolutions, and the establishment of interministerial bodies dedicated to digital transformation.[55] The National Council of Justice has assumed a leading role by issuing its own rules governing the use of AI within the Judiciary, signaling a trend toward sector-specific regulation. Meanwhile, Bill No. 2,338/2023 remains under legislative consideration, meaning that the current model continues to rely predominantly on executive guidelines and programmatic instruments.

4.2.2. Brazil's multilevel AI governance

Brazil illustrates the most layered approach to AI regulation in the region, combining federal, state, municipal, judicial, electoral, and criminal norms. At the federal level, the centerpiece is Bill No. 2,338/2023[56], currently pending in the National Congress, which adopts a risk-based architecture distinguishing "*excessive risk*" (prohibited uses) from "*high risk*" systems, and establishes civil liability rules and administrative sanctions. Bill 2,338/2023 also contains provisions on copyright and related rights in the context of AI development. [57]Although not yet enacted, the bill is complemented by a dense web of enforceable norms that, while not AI-specific, form its

---

53 *Ibid.*

54 *Ibid.*

55 *Ibid*.

56 BRASIL. Senado Federal. *Projeto de Lei nº 2.338, de 2023*. Dispõe sobre o uso da inteligência artificial. Brasília, DF: Senado Federal, 2023. Disponível em: https://legis.senado.leg.br/sdleggetter/documento?dm=9915448&ts=1774551291270&disposition=inline. Acesso em: 02 ago. 2026.

57 *Ibid*.

supporting framework: the General Data Protection Law, (13,709/2018[58]) whose Article 20 guarantees the right to review decisions made solely on the basis of automated processing; the Digital Government Law[59]; and the recent "Digital ECA" (Child and Adolescent Statute), Law 15,211/2025[60], which mandates periodic review of AI tools embedded in parental-control features and bans emotional-analysis and profiling techniques in advertising directed at minors.

4.2.3. Subnational lawmaking

A distinctive feature of the Brazilian case is the emergence of subnational AI legislation. Alagoas[61] enacted the country's first AI law (Law 9,095/2023), a concise statute setting out concepts and thirteen guiding principles, from human oversight and non-discrimination to due process, traceability, and accountability. Notably, the Alagoas statute extends diversity beyond outcomes to the composition of the teams that develop AI systems themselves, which should reflect gender, racial, ethnic, and age diversity, as well as the inclusion of persons with disabilities.[62] Paraná followed with Law 22,343/2025, built on five principles (including transparency, accountabili-

---

58 BRASIL. *Lei nº 13.709, de 14 de agosto de 2018*. Lei Geral de Proteção de Dados Pessoais (LGPD). Brasília, DF: Presidência da República, 2018. Disponível em: https://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/l13709.htm. Acesso em: 02 ago. 2026.

59 BRASIL. *Lei nº 14.129, de 29 de março de 2021*. Dispõe sobre princípios, regras e instrumentos para o Governo Digital e para o aumento da eficiência pública. Brasília, DF: Presidência da República, 2021. Disponível em: https://www.planalto.gov.br/ccivil_03/_ato2019-2022/2021/lei/l14129.htm. Acesso em 02 ago. 2026.

60 BRASIL. *Lei nº 15.211, de 17 de setembro de 2025*. Dispõe sobre a proteção de crianças e adolescentes em ambientes digitais (ECA Digital). Brasília, DF: Presidência da República, 2025. Disponível em: https://www.planalto.gov.br/ccivil_03/_ato2023-2026/2025/lei/l15211.htm. Acesso em: 02 ago. 2026.

61 ALAGOAS. *Lei nº 9.095, de 11 de dezembro de 2023*. ESTABELECE OS PRINCÍPIOS E DIRETRIZES PARA O USO DA INTELIGÊNCIA ARTIFICIAL, NO ÂMBITO DA ADMINISTRAÇÃO PÚBLICA ESTADUAL, E DÁ OUTRAS PROVIDÊNCIAS. Maceió: Assembleia Legislativa do Estado de Alagoas, 2023. Disponível em: https://sapl.al.al.leg.br/media/sapl/public/normajuridica/2023/2799/lei_no_9.095_de_11_de_dezembro_de_2023_.pdf. Acesso em: 06 ago. 2026.

62 *Ibid*.

ty, and non-discrimination) and expressly prohibiting the use of AI to invade privacy or to conduct unauthorized investigations.[63] Goiás went further, enacting a detailed Complementary Law 205/2025 that institutes a State Policy for AI Innovation and, notably, a permanent regulatory sandbox designed to reduce regulatory barriers and provide legal certainty for innovative AI initiatives, with an explicit commitment to open-source models and interoperable standards, alongside a broad set of guiding principles. [64]At the municipal level, the City of Rio de Janeiro established a program (Law 8,926/2025[65]) on the responsible use of AI, echoing the same thirteen-principle catalogue of the Alagoas AI Law[66] but adding six original guidelines, among them the promotion of public and private investment in AI research and development. A noteworthy feature of the municipal statute[67] is its interpretive clause (Article 3), which directs that the law be read not only in light of its stated foundations, objectives, and principles, but also with regard to the relevance of AI for innovation, competitiveness, and sustainable and inclusive economic growth, as well as for hu-

63 PARANÁ. *Lei nº 22.343/2025, de 04 de abril de 2025*. Dispõe sobre a regulamentação do uso da Inteligência Artificial no Estado do Paraná. Curitiba: Assembleia Legislativa do Estado do Paraná, 2025. Disponível em: https://www.legisweb.com.br/legislacao/?id=476131. Acesso em: 06 ago. 2026.

64 GOIÁS. *Lei Complementar nº 205, de* 19 de maio de 2025. Institui a Política Estadual de Fomento à Inovação em Inteligência Artificial no Estado de Goiás. Goiânia: Casa Civil do Estado de Goiás, 2025. Disponível em: https://legisla.casacivil.go.gov.br/api/v2/pesquisa/legislacoes/110694/pdf. Acesso em:

65 RIO DE JANEIRO (Município). *Lei nº 8.926, de 2025*. Institui o Programa sobre o Uso Responsável da Inteligência Artificial no âmbito do Município e dá outras providências. Rio de Janeiro: Câmara Municipal do Rio de Janeiro, 2025. Disponível em: https://leis.org/municipais/rj/rio-de-janeiro/lei/lei-ordinaria/2025/8926/. Acesso em: 06 ago. 2026.

66 ALAGOAS. *Lei nº 9.095, de 11 de dezembro de 2023*. ESTABELECE OS PRINCÍPIOS E DIRETRIZES PARA O USO DA INTELIGÊNCIA ARTIFICIAL, NO ÂMBITO DA ADMINISTRAÇÃO PÚBLICA ESTADUAL, E DÁ OUTRAS PROVIDÊNCIAS. Maceió: Assembleia Legislativa do Estado de Alagoas, 2023. Disponível em: https://sapl.al.al.leg.br/media/sapl/public/normajuridica/2023/2799/lei_no_9.095_de_11_de_dezembro_de_2023_.pdf. Acesso em: 06 ago. 2026.

67 RIO DE JANEIRO (Município). *Lei nº 8.926, de 2025*. Institui o Programa sobre o Uso Responsável da Inteligência Artificial no âmbito do Município e dá outras providências. Rio de Janeiro: Câmara Municipal do Rio de Janeiro, 2025. Disponível em: https://leis.org/municipais/rj/rio-de-janeiro/lei/lei-ordinaria/2025/8926/. Acesso em: 06 ago. 2026.

man and social development, always with a view to reducing bureaucracy and streamlining processes within the municipality's public sector.

4.2.4. Sectoral governance: judiciary, elections, and criminal law

Beyond legislation, Brazil has developed strong sector-specific governance. In the Judiciary, CNJ Resolution No. 615/2025[68] provides a binding framework for the development and use of AI by the courts, structured around principles of transparency, explainability, contestability, and effective human oversight, and incorporating its own risk-classification annex. A comparison between Bill 2,338/2023 and CNJ Resolution 615/2025 reveals two distinct techniques for handling risk. The Bill[69] fixes the risk taxonomy in the legislative text, distinguishing "*excessive risk*" (prohibited uses) from "*high-risk*" systems. The Resolution[70] (Article 9, §2), by contrast, sets out the risk-based structure but delegates the substantive criteria to the National Committee on AI of the Judiciary, to be defined after consultation with courts, developers, and civil society. The Brazilian framework thus combines a closed, legislatively fixed taxonomy with an open model whose criteria are set by a specialized body.

In the electoral domain, The Superior Electoral Court, through Resolution 23,755/2026, amended Resolution 23,610/2019[71], to require explicit labeling of

---

[68] BRASIL. Conselho Nacional de Justiça. *Resolução nº 615, de 11 de março de 2025*. Estabelece diretrizes para o desenvolvimento, utilização e governança de soluções desenvolvidas com recursos de inteligência artificial no Poder Judiciário. Brasília, DF: CNJ, 2025. Disponível em: https://atos.cnj.jus.br/files/original1555302025031467d4517244566.pdf. Acesso em: 06 ago. 2026.

[69] BRASIL. Senado Federal. *Projeto de Lei nº 2.338, de 2023.* Dispõe sobre o uso da inteligência artificial. Brasília, DF: Senado Federal, 2023. Disponível em: https://legis.senado.leg.br/sdleggetter/documento?dm=9915448&ts=1774551291270&disposition=inline. Acesso em: 06 ago. 2026.

[70] BRASIL. Conselho Nacional de Justiça. *Resolução nº 615, de 11 de março de 2025*. Estabelece diretrizes para o desenvolvimento, utilização e governança de soluções desenvolvidas com recursos de inteligência artificial no Poder Judiciário. Brasília, DF: CNJ, 2025. Disponível em: https://atos.cnj.jus.br/files/original1555302025031467d4517244566.pdf. Acesso em: 06 ago. 2026.

[71] BRASIL. Tribunal Superior Eleitoral. *Resolução nº 23.670, de 14 de dezembro de 2021.* Dispõe

synthetic media generated by AI in campaign advertising and to prohibit deepfakes reproducing the image or voice of candidates and public figures, to require explicit labeling of synthetic media generated by AI in campaign advertising and to prohibit deepfakes reproducing the image or voice of candidates and public figures. Finally, a 2025 amendment to the Criminal Code [72]introduced an aggravating factor increasing penalties by half where an offense is committed through AI or other technology that alters the victim's image or voice.

4.2.5 A broadened conception of AI sovereignty

The Brazilian strategy is distinctive in coupling regulation with an explicit concern for technological and informational sovereignty. In the PBIA[73], sovereignty means not only normative autonomy, the capacity to regulate the technology, but also the construction of the country's own infrastructural and computational capabilities: a "*sovereign cloud*" for critical and sensitive data stored on national territory and managed by public bodies, a national software stack, curated national datasets, and Portuguese-language large language models trained on domestic data. The concern extends to regulatory capacity itself, as the Plan envisions a national research network aimed at reducing external dependence in data and AI governance and at producing regulatory knowledge suited to the Brazilian context. [74]The Brazilian case

sobre as federações de partidos políticos. Brasília, DF: TSE, 2021. Disponível em: https://www.tse.jus.br/legislacao/compilada/res/2021/resolucao-no-23-670-de-14-de-dezembro-de-2021. Acesso em: 06 ago. 2026.

72 BRASIL. *Lei nº 15.123, de 24 de abril de 2025.* Altera o art. 147-B do Decreto-Lei nº 2.848, de 7 de dezembro de 1940 (Código Penal), para estabelecer causa de aumento de pena no crime de violência psicológica contra a mulher quando praticado com o uso de inteligência artificial ou de qualquer outro recurso tecnológico que altere imagem ou som da vítima. Brasília, DF: Presidência da República, 2025. Disponível em: https://www.planalto.gov.br/ccivil_03/_ato2023-2026/2025/lei/l15123.htm. Acesso em 06 ago. 2026.

73 BRASIL. Ministério da Ciência, Tecnologia e Inovação. *Plano Brasileiro de Inteligência Artificial (PBIA): IA para o bem de todos.* Brasília, DF: MCTI, 2024. Disponível em: https://www.gov.br/mcti/pt-br/acompanhe-o-mcti/transformacaodigital/plano-brasileiro-de-inteligencia-artificial-pbia-_vf.pdf. Acesso em 02 ago. 2026.

74 *Ibid*.

thus points to a broader notion of AI sovereignty, one that spans not only the autonomy to regulate the technology, but also the capacity to develop and govern it.

4.3 CHILE

Chile stands out, first of all, for having been the first country in the world to give neurorights constitutional standing, a pioneering step in shielding mental activity from the reach of emerging technologies. Its wider regulatory infrastructure is also comparatively developed[75]. The Cybersecurity Framework Law of 2024 requires systems to be conceived from the outset with protection built in rather than added later, measures cyber-risk by weighing how likely an incident is against the harm it would cause, and sets up a National Cybersecurity Agency as a dedicated technical body.[76] Alongside it, a new Data Protection Law, approved in 2024 and effective from 1 December 2026, implicitly addresses explainability through a right to an explanation, provides for human intervention, and requires that the logic behind any automated decision be disclosed.[77]

At the level of strategy, the National AI Policy updated in 2024, framed in line with the UNESCO Recommendation, moves beyond its 2021 version by adding new priorities: equipping the state to use AI ethically, taking an active part in international governance debates, harnessing renewable energy and AI against the climate crisis, and revisiting labour rules to safeguard workers. [78]The accompanying 2024 AI bill follows a risk-based logic and reaches, among other things, systems designed to

---

[75] UNESCO. *Chile: pioneering the protection of neurorights*. The UNESCO Courier, Paris, 2022. Disponível em: https://courier.unesco.org/en/articles/chile-pioneering-protection-neurorights. Acesso em: 6 ago. 2026.

[76] CHILE. *Ley 21.663, Ley Marco de Ciberseguridad*. Santiago: Biblioteca del Congreso Nacional de Chile, 2024. Disponível em: http://bcn.cl/leychile/navegar?idNorma=1202434. Acesso em: 6 ago. 2026.

[77] CHILE. *Ley 21.719, que regula la protección y el tratamiento de los datos personales y crea la Agencia de Protección de Datos Personales*. Santiago: Biblioteca del Congreso Nacional de Chile, 2024. Disponível em: https://www.bcn.cl/leychile/navegar?idNorma=1209272. Acesso em: 6 ago. 2026

[78] OECD.AI. *National AI policies: Chile*. Paris: OECD, 2026. Disponível em: https://oecd.ai/en/dashboards/national/chile. Acesso em: 6 ago. 2026

recognise emotions. [79]The bill provides for supervised testing spaces, a standing multisectoral advisory council on AI comprising government, regulators, academia, industry, and civil society, administrative penalties graded by severity, and judicial review of AI-related harm on a fault-based standard.[80]

4.4 COLOMBIA

Colombia has advanced its AI framework primarily through policy instruments rather than a single enacted statute, building on the data-protection regime of Law 1581 of 2012[81]. Colombia's AI agenda is also embedded in broader digital-policy instruments, such as the National Digital Strategy 2023–2026[82], which identifies the strengthening of AI capabilities as a strategic priority. Its most comprehensive document, the National AI Policy (CONPES 4144[83]), sets the goal of building capacities for the ethical and sustainable adoption of AI to drive the country's transformation by 2030, organized around six objectives spanning governance, infrastructure, research, talent, risk mitigation, and cross-sector adoption.

Complementing this, Bill 422 of 2025 seeks to give AI regulation a legislative footing. A distinctive feature is its detailed treatment of regulatory sandboxes: it creates a National Technical Committee, coordinated by the Ministry of Science,

---

79 CHILE. Cámara de Diputadas y Diputados. *Proyecto de ley que regula los sistemas de inteligencia artificial*: boletín n. 16.821-19. Santiago: Cámara de Diputadas y Diputados. Disponível em: https://www.camara.cl/legislacion/proyectosdeley/tramitacion.aspx?prmID=17429&prmBOLETIN=16821-19. Acesso em: 6 ago. 2026.

80 *Ibid*.

81 COLÔMBIA. *Ley 1581 de 2012.* Por la cual se dictan disposiciones generales para la protección de datos personales. Bogotá: Congreso de la República, 2012. Disponível em: https://www.funcionpublica.gov.co/eva/gestornormativo/norma.php?i=49981. Acesso em: 06 ago. 2026.

82 COLÔMBIA. *Estrategia Nacional Digital de Colombia 2023–2026*. Bogotá: Colombia Government, 2023. Disponível em: https://siteal.iiep.unesco.org/sites/default/files/sit_accion_files/estrategia_nacional_digital_de_colombia_2023-2026.pdf. Acesso em: 06 ago. 2026.

83 COLÔMBIA. Departamento Nacional de Planeación. *Documento CONPES 4144: Política Nacional de Inteligencia Artificial*. Bogotá: DNP, 2025. Disponível em: https://siteal.iiep.unesco.org/sites/default/files/sit_accion_files/documento_conpes_4144_-_politica_nacional_de_inteligencia_artificial.pdf. Acesso em: 06 ago. 2026.

Technology and Innovation, to approve a binding plan for each authorized sandbox, alongside provisions on administrative sanctions.[84]

Colombia's Roadmap for the Development and Application of AI[85] grounds the framework in the ethical principles of UNESCO, the UN, and the OECD, and develops a notably detailed conception of algorithmic transparency, distinguishing reactive transparency, where public bodies disclose information on request, from active transparency, where it is provided permanently or periodically. Transparency, in this view, is not merely a monitoring tool but a condition enabling data protection, privacy, and access to information.

Colombia has also been a pioneer in the judicial domain. Although no national legislation yet governs AI in court proceedings, in December 2024 the Superior Council of the Judiciary adopted Guidelines for the Responsible and Safe Use of Generative AI in the Judicial Branch, structured around fifteen principles , making Colombia the first country to implement the UNESCO recommendation on AI ethics within its judiciary[86].

4.5 COSTA RICA

Notably broad and concise in scope, Bill 23,771 of 2023 is reported to have been drafted with the help of ChatGPT.[87] It establishes five guiding principles and, among its provisions, obliges AI systems whose automated decisions affect individuals to

---

84 COLÔMBIA. Ministerio de Ciencia, Tecnología e Innovación. *Proyecto de Ley nº 422 de 2025*. Bogotá: MinCiencias, 2025. Disponível em: https://minciencias.gov.co/sites/default/files/upload/noticias/pl_ia_finalizado.pdf. Acesso em: 06 ago. 2026.

85 COLÔMBIA. Ministerio de Ciencia, Tecnología e Innovación. *Hoja de ruta para el desarrollo y aplicación de la inteligencia artificial en Colombia.* Bogotá: MinCiencias, 2024. Disponível em: https://minciencias.gov.co/sites/default/files/upload/noticias/hoja_de_ruta_adopcion_etica_y_sostenible_de_inteligencia_artificial_colombia_0.pdf. Acesso em: 006 ago. 2026.

86 TECH AND JUSTICE. *Colombia*. Oxford: Blavatnik School of Government, University of Oxford, 2026. Disponível em: https://www.techandjustice.bsg.ox.ac.uk/research/colombia. Acesso em: 06 ago. 2026.

87 COSTA RICA. Asamblea Legislativa. *Proyecto de Ley nº 23.771*. San José: Asamblea Legislativa de la República de Costa Rica, 2023. Disponível em: : https://delfino.cr/asamblea/proyecto/23771. Acesso em: 06 ago. 2026.

explain, in clear and understandable terms, the reasons and criteria on which those decisions rest.[88]

Beyond this legislative initiative, Costa Rica's National AI Strategy is built on a set of cross-cutting principles, among which the promotion of research, development, and innovation (R&D&I) stands out, conceived so that technological progress benefits society as a whole on an equitable basis, in tandem with the protection of human life.[89] In line with the strategy's risk-oriented outlook, four levels of risk are defined, unacceptable, high, medium, and low, drawing on the European Union's AI Act but adjusting its categories to Costa Rica's own circumstances and strategic priorities. [90]

At the institutional level, governance centers on the National Agency for Digital Government (NADG), which already drives the digital transformation of public institutions and, under the strategy, assumes an even more decisive role in risk management, in the delivery of AI-related services, and as the body responsible for interoperability, treated here as a key condition for managing the risks of AI deployment. [91] The strategy further contemplates the enactment of "*specific regulations, decrees, and guidelines*" to steer the development and use of AI.[92] Finally, the strategy is organized around seven axes, one of which, territorial articulation, seeks to strengthen regional integration and spur economic development through the uptake of AI across all parts of the country, with particular attention to areas beyond the Greater Metropolitan Area (GAM).[93]

4.6 ECUADOR

Ecuador's strategy frames artificial intelligence as a means of advancing economic

---

88 *Ibid*.

89 COSTA RICA. Ministerio de Ciencia, Innovación, Tecnología y Telecomunicaciones. *National Artificial Intelligence Strategy of Costa Rica*. San José: MICITT, 2024. Disponível em: https://www.micitt.go.cr/sites/default/files/2025-02/National%20Artificial%20Intelligence%20Strategy%20of%20Costa%20Rica.pdf. Acesso em: 06 ago. 2026.

90 *Ibid*.

91 *Ibid*.

92 *Ibid*.

93 *Ibid*.

development, social inclusion, and environmental sustainability while respecting citizens' fundamental rights, in what it describes as a genuine democratisation of technology.[94] The document frames AI as both an opportunity and a challenge, one to be approached on ethical and responsible terms, with attention to the country's cultural and linguistic diversity as well as to existing socioeconomic gaps.[95] Covering the period from 2025 to 2029, the strategy emerged from a participatory process involving experts, academics, and citizens. Its formulation also incorporated international references, including guidance on AI public policy issued by the Development Bank of Latin America and the Caribbean (CAF)[96]. A central objective is to distribute the benefits of AI across Ecuadorian society, with particular attention to rural areas, Indigenous communities, and small businesses.[97] In accordance with the SCE Code of Ethics, its Guide on the Use of AI Tools, and the applicable Ecuadorian legal framework, the relevant principles include transparency and explainability, security and cyber-resilience, which also introduces a cybersecurity dimension into AI governance, and the promotion of a human-centred and participatory AI culture.[98]

The 2024 Draft Organic Law on the Regulation and Promotion of Artificial Intelligence seeks to provide a framework for the development and use of AI in Ecuador that combines the protection of democracy, human rights, and the rule of law with innovation and the social benefits associated with these technologies.[99] The proposal

---

94 EQUADOR. Asamblea Nacional. *Proyecto de Ley Orgánica de Regulación y Promoción de la Inteligencia Artificial en Ecuador*. Quito, 2024. Disponível em: https://www.asambleanacional.gob.ec/sites/default/files/private/asambleanacional/filesasambleanacionalnameuid-19130/2192.%20Proyecto%20de%20Ley%20Organica%20de%20Regulacion%20y%20Promocion%20de%20la%20Inteligencia%20Artificial%20en%20Ecuador%20-pnunez/pp%20-%20proyecto%20de%20ley%20450889-nunez.pdf. Acesso em: 6 ago. 2026.

95 *Ibid*.

96 *Ibid*.

97 *Ibid*.

98 EQUADOR. SUPERINTENDENCIA DE COMPETENCIA ECONÓMICA. *Decálogo ético para el uso de la inteligencia artificial*. 2026. Disponível em: https://oecd.ai/en/dashboards/policy-initiatives/ethical-decalogue-for-the-use-of-artificial-intelligence-ai. Acesso em: 6 ago. 2026.

99 EQUADOR. Asamblea Nacional. *Proyecto de Ley Orgánica de Regulación y Promoción de la Inteligencia Artificial en Ecuador*. Quito, 2024. Disponível em: https://www.asambleanacional.gob.ec/sites/default/files/private/asambleanacional/filesasambleanacionalnameuid-

adopts a four-tier risk classification, low, moderate, high, and extreme, and extends its regulatory scope to copyright and related rights in the context of AI. It also provides for regulatory sandboxes and establishes a system of administrative sanctions. [100]

4.7 MEXICO

Mexico has not yet adopted a comprehensive national AI policy, the most recent instrument of national scope remaining the 2018 National Artificial Intelligence Strategy[101], drafted before the generative-AI cycle. The gap is notable given the country's regional weight: Mexico and Brazil together account for some ninety-five per cent of AI-related patents filed in Latin America, although Mexico ranks only sixth in the region by number of researchers working specifically on AI.[102] In the absence of a general framework, the legislature has proceeded through targeted amendments to existing statutes, expressly postponing a comprehensive law.

A preliminary obstacle is one of legislative competence. A pending constitutional initiative would amend fraction XVII of Article 73[103], the provision setting out the

19130/2192.%20Proyecto%20de%20Ley%20Organica%20de%20Regulacion%20y%20Promocion%20de%20la%20Inteligencia%20Artificial%20en%20Ecuador%20-pnunez/pp%20-%20proyecto%20de%20ley%20450889-nunez.pdf. Acesso em: 6 ago. 2026.

[100] *Ibid*.

[101] LÓPEZ MARTÍNEZ, Víctor José. Hacia la ley mexicana de la inteligencia artificial y sus derivados: buques autónomos, drones y robots. *El Financiero*, Ciudad de México, 3 ago. 2026. Opinión. Disponível em: https://www.elfinanciero.com.mx/opinion/victor-jose-lopez-martinez/2026/08/03/hacia-la-ley-mexicana-de-la-inteligencia-artificial-y-sus-derivados-buques-autonomos-drones-y-robots/. Acesso em: 6 ago. 2026.

[102] UNESCO. *Ethics of artificial intelligence: Mexico*. Paris: UNESCO, 2026. Disponível em: https://www.unesco.org/ethics-ai/en/mexico. Acesso em: 6 ago. 2026.

[103] MÉXICO. *Constitución Política de los Estados Unidos Mexicanos.* Ciudad de México: Suprema Corte de Justicia de la Nación, [2026]. Disponível em: https://www.scjn.gob.mx/sites/default/files/cpeum/documento/cpeum.pdf. Acesso em: 6 ago. 2026; DIPUTADOS de Morena anuncian iniciativa para facultar al Congreso legislar sobre la IA. *La Jornada,* Ciudad de México, 6 ago. 2026. Política. Disponível em: https://www.jornada.com.mx/noticia/2026/08/08/politica/diputados-de-morena-anuncia-iniciativa-para-facultar-al-congreso-legislar-sobre-la-ia. Acesso em: 6 ago. 2026.

powers of Congress, so that its authority to enact legislation on general means of communication expressly encompasses artificial intelligence.

The reforms already adopted converge on a single concern, the unauthorised synthetic reproduction of a person's identity. A reform to the Federal Copyright Law [104] brings personal image and voice within the scope of protection against AI-generated outputs, providing that a performer's voice and image, including the characters they portray, may be used or published only with express consent. In addition, Article 305 of the Federal Labour Law[105] was amended, requiring artists, performers and executants contracts to detail AI-specific terms and to state the remuneration payable for any AI processing of the worker's image or voice.

Underlying both reforms is the 2025 Federal Law on the Protection of Personal Data[106], which supplies the baseline rules on the processing of personal information and therefore conditions the training and deployment of AI systems generally. Taken together, the three instruments illustrate the Mexican method: rather than defining artificial intelligence as an autonomous regulatory object, the legislature has distributed its treatment across the branches of law where the harms first materialised, authorship, employment and data protection.

A further legislative proposal would amend the Federal Criminal Code[107] to create

---

[104] MÉXICO. *Ley Federal del Derecho de Autor*, de 24 de diciembre de 1996. Última reforma publicada DOF 14-05-2026. Ciudad de México: Cámara de Diputados, 2026. Disponível em: https://www.diputados.gob.mx/LeyesBiblio/pdf/LFDA.pdf. Acesso em: 6 ago. 2026; CADENA POLÍTICA. *México: reforma a la Ley Federal del Trabajo y a la Ley Federal del Derecho de Autor sobre inteligencia artificial y artistas*. Cadena Política, 15 maio 2026. Disponível em: https://cadenapolitica.com/2026/05/15/mexico-reforma-ley-federal-trabajo-derecho-autor-ia-artistas. Acesso em: 6 ago. 2026;

[105] Ibid; MÉXICO. *Ley Federal del Trabajo*, de 1º de abril de 1970. Última reforma publicada DOF 14-05-2026. Ciudad de México: Cámara de Diputados, 2026. Disponível em: https://www.diputados.gob.mx/LeyesBiblio/pdf/LFT.pdf. Acesso em: 6 ago. 2026.

[106] MÉXICO. *Ley Federal de Protección de Datos Personales en Posesión de los Particulares*. Nueva ley publicada en el Diario Oficial de la Federación el 20 de marzo de 2025. Última reforma publicada DOF 14-11-2025. Disponível em: https://dgesui.ses.sep.gob.mx/sites/default/files/2026-01/LFPDPPP.pdf. Acesso em: 6 ago. 2026.

[107] MÉXICO. Secretaría de Gobernación. Sistema de Información Legislativa. *INICIATIVA QUE REFORMA EL ARTÍCULO 199 OCTIES DEL CÓDIGO PENAL FEDERAL, EN MATERIA DE*

the offence of violation of sexual intimacy, expressly covering intimate images, audio or video produced through technological tools, algorithms or artificial intelligence systems without the subject's consent.

A draft General Law on Artificial Intelligence[108] is nonetheless in the legislative process. Two of its guiding principles are particularly indicative of its orientation. The first requires that understanding who and what will be affected must precede any decision to automate, converting impact assessment into a precondition of deployment rather than a subsequent verification.[109] The second holds that whatever cannot be explained should not be automated, making explainability a criterion of admissibility rather than a mere transparency duty, a demanding standard, given that it excludes from lawful use any system whose reasoning cannot be reconstructed.[110]

The bill also reaches neurorights. Article 18 recognises five, "*personal identity and psychological continuity*"; "*mental privacy*"; "*cognitive integrity and protection against non-consented interference*"; "*autonomy of the will*"; and "eq*uity and freedom from undue neurotechnological manipulation*."[111]

The statute is not conceived as a self-contained regime. Article 9 provides that it is to be interpreted and applied in harmony with the Federal Constitution and, on an illustrative rather than exhaustive basis, with a list of some forty-five federal laws, among them the Federal Copyright Law and the Federal Labour Law discussed above.[112] Article 10 extends the same logic to the international plane, requiring application consistent with the treaties to which Mexico is a party, and in particular

---

VIOLENCIA DIGITAL, PROTECCIÓN A LA INTIMIDAD Y SANCIÓN DEL USO INDEBIDO DE TECNOLOGÍAS E INTELIGENCIA ARTIFICIAL: asunto n. 5.015.968. Ciudad de México, 17 fev. 2026. Disponível em: https://sil.gobernacion.gob.mx/Archivos/Documentos/2026/02/asun_5015968_20260217_1771373777.pdf. Acesso em: 6 ago. 2026.

[108] MÉXICO. Senado de la República. Comisión de Análisis, Seguimiento y Evaluación sobre la Aplicación y Desarrollo de la Inteligencia Artificial en México. *Iniciativa con proyecto de decreto por el que se expide la Ley General para Regular y Fomentar el Uso de la Inteligencia Artificial (IA) en México*. Ciudad de México: LXVI Legislatura del Congreso de la Unión, mar. 2026. 134 p.

[109] *Ibid*.

[110] *Ibid*.

[111] *Ibid*.

[112] *Ibid*.

with UNESCO instruments, OECD recommendations and the Mexico–United States–Canada Agreement (T-MEC/USMCA), provided that such consistency remains compatible with technological sovereignty, national security and the Mexican public interest.

4.8 PANAMA

Legislative activity in Panama is centred on Bill 588, which establishes the *"adaptive governance of artificial intelligence*" and embraces regulatory experimentation in accordance with the UNESCO Recommendation. [113] Following its approval in first debate by the National Assembly in April 2026, the bill remains subject to further legislative deliberation.[114] Its adaptive character was emphasized when the proposal was presented as establishing a form of governance capable of evolving through public policy laboratories, while also seeking to foster foreign investment and technological development.[115] On the institutional front, Executive Decree 36 of May 2026 established the National Commission for Critical and Emerging Technologies and assigned the chairmanship of its AI Subcommission to SENACYT.[116] These developments rest on an earlier foundation that predates the problems now at issue: Law 81 of 2019 recognises ARCO rights and protects individuals from decisions based exclusively on automated processing.[117] Its coverage, however, does not extend to "*algorithmic profiling, large-scale biometric processing, or synthetic data*". [118]

The policy framework surrounding this legislation is set out in the National Stra-

---

[113] UNESCO. Ethics of artificial intelligence: Panama. Paris: UNESCO, 2026. Disponível em: https://www.unesco.org/ethics-ai/en/panama. Acesso em: 6 ago. 2026.

[114] PANAMÁ. Asamblea Nacional. Panamá a la vanguardia de la tecnología con IA, según proyecto de ley aprobado en primer debate. Ciudad de Panamá, 2026. Disponível em: https://www.asamblea.gob.pa/Noticias/Noticias/PANAMA-A-LA-VANGUARDIA-DE-LA-TECNOLOGIA-CON-IA-SEGUN-PROYECTO-DE-LEY-APROBADO-EN-PRIMER-DEBATE. Acesso em: 6 ago. 2026.

[115] *Ibid*.

[116] UNESCO. *Ethics of artificial intelligence:* Panama. Paris: UNESCO, 2026. Disponível em: https://www.unesco.org/ethics-ai/en/panama. Acesso em: 6 ago. 2026.

[117] *Ibid*.

[118] *Ibid*.

tegic Plan for Science and Technology (PENCYT) 2025–2029, structured around six cross-cutting development vectors covering areas regarded as strategic to Panama including productive transformation, health, social and institutional development. [119]The purpose of this architecture is to draw on knowledge generated in fields of particular relevance, including renewable energy, Industry 5.0 and emerging technologies such as artificial intelligence, the Internet of Things, blockchain and cybersecurity.[120] Artificial intelligence accordingly enters Panamanian policy not as an isolated regulatory object but as one component of a broader technological agenda. [121]

That agenda was consolidated in July 2026 with the launch of the National Agenda for Advanced Technologies, presented against the observation that artificial intelligence, microelectronics and semiconductors have become foundations of the global economy.[122] The Agenda also connects Bill 588 to this broader policy orientation, presenting the proposed framework as a modern, risk-based and experimental approach that combines the protection of citizens' rights with greater certainty for investment.[123] The same orientation governs AI 2036, Panama's National Artificial Intelligence Strategy, two of whose principles are especially indicative.[124] First, the strategy frames AI sovereignty in terms of Panama's ability to build domestic capabi-

---

[119] PANAMÁ. Secretaría Nacional de Ciencia, Tecnología e Innovación. Plan Estratégico Nacional de Ciencia, Tecnología e Innovación (PENCYT) 2025-2029. Ciudad de Panamá: SENACYT, 2025. Disponível em: https://www.senacyt.gob.pa/wp-content/uploads/2025/05/PENCYT-2025-2029.pdf. Acesso em: 6 ago. 2026

[120] *Ibid*.

[121] *Ibid*.

[122] PANAMÁ. Secretaría Nacional de Ciencia, Tecnología e Innovación. Panamá lanza la Agenda Nacional de Tecnologías Avanzadas en presencia del presidente Mulino. Ciudad de Panamá, 30 jul. 2026. Disponível em: https://www.senacyt.gob.pa/en/panama-lanza-la-agenda-nacional-de-tecnologias-avanzadas-en-presencia-del-presidente-mulino/. Acesso em: 6 ago. 2026.

[123] PANAMÁ. Asamblea Nacional. Panamá a la vanguardia de la tecnología con IA, según proyecto de ley aprobado en primer debate. Ciudad de Panamá, 2026. Disponível em: https://www.asamblea.gob.pa/Noticias/Noticias/PANAMA-A-LA-VANGUARDIA-DE-LA-TECNOLOGIA-CON-IA-SEGUN-PROYECTO-DE-LEY-APROBADO-EN-PRIMER-DEBATE. Acesso em: 6 ago. 2026.

[124] PANAMÁ. *AI 2036: National Artificial Intelligence Strategy of Panama.* Ciudad de Panamá, 2026. Disponível em: https://aipanama.gob.pa/. Acesso em: 6 ago. 2026.

lities for understanding, governing, and operating AI systems of strategic importance to national security and the economy, while limiting reliance on opaque foreign systems in "*essential government functions*".[125] Second, it embraces an experimental and agile approach to governance: in response to rapid technological change, "*flexible sandboxes*" and "*iterative policymaking*" are preferred to fixed regulatory arrangements.[126] Two strategic actions give these principles operational content an AI Risk Observatory, to be established by SENACYT, will track international trends in AI safety and document domestic cases of algorithmic harm, reporting annually to legislators and regulators so that policy can evolve alongside technological developments.[127]

4.9 PERU

Peru stands apart from much of the region in having moved beyond the drafting stage. it enacted binding AI legislation in 2023 and issued implementing regulations two years later. Law 31,814 of 2023 is a short statute, yet the framework it set in motion has since been given considerable regulatory detail.[128] Its stated purpose is to encourage the use of artificial intelligence within the country's broader digital transformation, placing the human person and respect for human rights at the centre, so that economic and social development advances in a setting that keeps such use "*ethical, sustainable, transparent, reproducible and accountable*". [129]Among the principles it lays down is the adoption of risk-based security standards, an early signal that Peru's model would calibrate obligations to the level of danger a given system presents. [130]

---

[125] *Ibid*.

[126] *Ibid*.

[127] *Ibid*.

[128] PERU. *Ley n. 31814, que promueve el uso de la inteligencia artificial en favor del desarrollo económico y social del país*. Lima, 2023. Disponível em: https://cdn.www.gob.pe/uploads/document/file/5038703/ley-que-promueve-el-uso-de-la-inteligencia-artificial-en-fav-ley-n-31814.pdf. Acesso em: 6 ago. 2026.

[129] *Ibid*.

[130] *Ibid*.

That calibration was developed in 2025, when the President of the Republic issued the regulation implementing Law 31,814. The regulation widens the set of principles established by the statute, adding, among others, respect for copyright and related rights, a concern that connects the Peruvian framework to the same anxieties over authorship and AI seen elsewhere in the region. [131]It also introduces a two-tier classification of AI uses, distinguishing improper use from high-risk use, and attaches distinct regulatory consequences to each. [132]

Underpinning these instruments is an older foundation. Legislative Decree 1412 of 2018, which established the Digital Government Law, set out the governance framework for digital government in Peru, covering "digital identity, digital services, digital architecture, interoperability, digital security and data, together with the legal regime for the cross-cutting use of digital technologies" across the public administration. [133] It is on this infrastructure that the country's later AI policy has been built.

The most recent piece is the National Artificial Intelligence Strategy 2026–2030, which aligns Peru with prevailing international standards by adopting the principles of the OECD Council Recommendation on Artificial Intelligence and its people-centred orientation.[134] The strategy is notable less for what it prescribes than for what it diagnoses. [135]Its central finding is that "*the mere existence of rules is not enough to secure a coherent and uniform adoption of artificial intelligence across the state*".[136] The state had adopted AI piecemeal, with investments that pulled in different direc-

---

[131] PERU. Decreto Supremo n. 115-2025-PCM, que aprueba el Reglamento de la Ley n. 31814, Ley que promueve el uso de la inteligencia artificial en favor del desarrollo económico y social del país. Diario Oficial El Peruano, Lima, 2025. Disponível em: https://busquedas.elperuano.pe/dispositivo/NL/2436426-1. Acesso em: 6 ago. 2026.

[132] Ibid.

[133] PERU. Decreto Legislativo n. 1412, que aprueba la Ley de Gobierno Digital. Diario Oficial El Peruano, Lima, 2018. Disponível em: https://cdn.www.gob.pe/uploads/document/file/353216/decreto-legislativo-que-aprueba-la-ley-de-gobierno-digital-decreto-legislativo-n-1412-1691026-1.pdf. Acesso em: 6 ago. 2026

[134] PERU. Estrategia Nacional de Inteligencia Artificial 2026-2030. Lima: PCM, 2026. Disponível em: https://cdn.www.gob.pe/uploads/document/file/10347430/8408055-estrategia-nacional-de-inteligencia-artificial-2026-2030.pdf. Acesso em: 6 ago. 2026.

[135] *Ibid*.

[136] *Ibid*.

tions, overlapping mandates, inefficient spending, corruption, and no dependable way to monitor or assess the technology's effects.[137] Against that diagnosis, the strategy proposes a shift from digital government towards intelligent government, organised around four axes, among them innovation and entrepreneurship grounded in artificial intelligence. [138]

This concern with uptake echoes an observation already made in the National Digital Transformation Policy to 2030, adopted in 2023, which noted that unfamiliarity with AI-based tools, or poor experiences with them, could harden into resistance, and that such resistance risked pushing citizens still further from the benefits these technologies might bring. [139]

Criminal law has also been adapted to address the misuse of AI. Law 32314 of 2025 introduced amendments to both the Criminal Code and the Cybercrime Law, treating the involvement of AI as an aggravating circumstance for several offences, such as defamation, fraud, child sexual abuse material, plagiarism, and copyright violations.[140] Enhanced penalties may therefore arise from the use of AI tools or technologies capable of manipulating images and voices, including deepfakes, when employed for defamatory impersonation, the creation of pornographic material involving manipulated images of minors, or economic fraud based on simulated identities or voices.[141]

---

[137] *Ibid*.

[138] *Ibid*.

[139] PERU. Presidencia del Consejo de Ministros. Política Nacional de Transformación Digital al 2030. Lima: PCM, 2023. Disponível em: https://cdn.www.gob.pe/uploads/document/file/4912655/Politica%20Nacional%20de%20Transformacion%20Digital%20al%202030_Resumen%20ejecutivo.pdf. Acesso em: 6 ago. 2026.

[140] PERU. *Ley n. 32314, que modifica el Código Penal y la Ley de Delitos Informáticos para incorporar el uso de inteligencia artificial como agravante en la comisión de diversos delitos*. Diario Oficial El Peruano, Lima, 2025. Disponível em: https://busquedas.elperuano.pe/dispositivo/NL/2394851-2. Acesso em: 6 ago. 2026.

[141] *Ibid*.

4.10 URUGUAY

Uruguay occupies an unusual position in the region. Alongside Argentina, it is one of only two Latin American countries whose data protection regime has been recognised as adequate under European standards.[142] The recognition matters well beyond privacy law, since it underwrites confidence in cross-border commerce involving AI systems that depend on data.[143] The domestic mandate for a national policy was established by Article 74 of Law 20.212, of November 2023, which charged Agesic, the agency responsible for electronic government and the information and knowledge society, with designing and developing a national data and artificial intelligence strategy anchored in international standards and applicable to both the public and private spheres.[144] Where the processing of personal data is concerned, the provision makes joint action with the Personal Data Regulatory and Control Unit (URCDP) mandatory rather than discretionary.[145]

The framework also draws explicitly on international commitments. The Montevideo Declaration reflects a regional commitment to an approach to AI grounded in "*human rights, fundamental freedoms, the rule of law, and democracy*".[146] It places human dignity and well-being at the centre of technological development, while seeking to reconcile innovation with inclusive and sustainable progress.[147] At the same time, the Declaration preserves a substantial margin of national autonomy,

---

142 UNESCO. *Ethics of artificial intelligence:* Uruguay. Paris: UNESCO, 2026. Global AI Ethics and Governance Observatory. Disponível em: https://www.unesco.org/ethics-ai/en/uruguay. Acesso em: 6 ago. 2026

143 *Ibid*.

144 URUGUAI. *Ley 20.212, de 6 de noviembre de 2023*: rendición de cuentas y balance de ejecución presupuestal, ejercicio 2022. Artículo 74. Montevidéu: IMPO, 2023. Disponível em: https://www.impo.com.uy/bases/leyes/20212-2023/74. Acesso em: 6 ago. 2026.

145 *Ibid*.

146 URUGUAI. Agencia de Gobierno Electrónico y Sociedad de la Información y del Conocimiento. Aprobación de la Declaración de Montevideo. Montevidéu: Agesic, 2024. Disponível em: https://www.gub.uy/agencia-gobierno-electronico-sociedad-informacion-conocimiento/comunicacion/noticias/aprobacion-declaracion-montevideo. Acesso em: 6 ago. 2026.

147 *Ibid*.

allowing each country to shape its policies in light of its own interests, priorities, circumstances, domestic legal framework, and international commitments.[148] Such autonomy, however, is presented as compatible with the need for regional and global governance, particularly in view of AI's implications across multiple domains.[149] That sovereign dimension is reinforced by the Declaration's insistence that the "*governance of artificial intelligence must take account of the particularities of the region, including its socio-economic and cultural contexts*".[150] The regional framework is further reinforced by the Council of Europe treaty on artificial intelligence, a legally binding international instrument aimed at ensuring that AI systems are developed and used consistently "*with human rights, democracy, and the rule of law*".[151]

Uruguay's National Artificial Intelligence Strategy 2024–2030 connects the country's adoption of AI with a broader agenda of social, economic, and institutional development.[152] Its objectives encompass responsible and secure technological development, the promotion of innovation, individual well-being, and improvements in public management, while incorporating ethical, critical, and creative dimensions into the use of AI.[153] These objectives are developed through three thematic axes: "*governance, AI capabilities, and sustainable development*".[154]

---

148 *Ibid*.

149 *Ibid*.

150 *Ibid*.

151 URUGUAI. Agencia de Gobierno Electrónico y Sociedad de la Información y del Conocimiento. *Tratado del Consejo de Europa sobre IA: Uruguay es el primer país de Latinoamérica en adherirse.* Montevidéu: Agesic, set. 2025. Disponível em: https://www.gub.uy/agencia-gobierno-electronico-sociedad-informacion-conocimiento/comunicacion/noticias/tratado-del-consejo-europa-sobre-ia-uruguay-es-primer-pais-latinoamerica. Acesso em: 6 ago. 2026.

152 URUGUAI. Agencia de Gobierno Electrónico y Sociedad de la Información y del Conocimiento. Estrategia Nacional de Inteligencia Artificial del Uruguay 2024-2030. Montevidéu: Agesic, 2024. Disponível em: https://www.gub.uy/agencia-gobierno-electronico-sociedad-informacion-conocimiento/comunicacion/publicaciones/estrategia-nacional-inteligencia-artificial-del-uruguay-2024-2030. Acesso em: 6 ago. 2026.

153 *Ibid*.

154 *Ibid*.

5. 2023 SANTIAGO DECLARATION

At the regional level, the 2023 Santiago Declaration, signed by nine of the countries examined, although not by Panama, frames AI as both an opportunity for development and a source of significant societal risks.[155] It calls for regional and multilateral coordination and assigns States a central role in shaping ethical governance capable of fostering the benefits of AI while addressing its implications for human rights, privacy, personal data, democracy, employment, and inequality.[156] The Declaration also stresses that AI development should reflect Latin American and Caribbean cultures, priorities, and capabilities, with meaningful regional participation in research and development.[157] In this context, the UNESCO Recommendation on the Ethics of Artificial Intelligence is identified as an important reference for building international consensus on the responsible development and use of AI.[158]

6. WHAT CAN BE LEARNED FROM THE LATIN AMERICAN REGULATORY TREND?

Despite differences in regulatory pace and national priorities, a shared pattern emerges across the countries examined. Initial institutional responses have largely taken the form of soft law and administrative measures, including national plans and guidelines. This choice has evident advantages, agility, adaptability, and the capacity for experimentation, but also limitations. Without a clear legal basis and enforcement mechanisms, guidelines may become merely performative statements. The literature on AI regulation, when addressing the trade-off between innovation and protection, suggests that the challenge lies in combining flexibility with instruments

---

155 CÚPULA MINISTERIAL E DE ALTAS AUTORIDADES DA AMÉRICA LATINA E DO CARIBE. *Declaración de Santiago: para promover una inteligencia artificial ética en América Latina y el Caribe*. Santiago de Chile, 23-24 out. 2023. Disponível em: https://minciencia.gob.cl/uploads/filer_public/40/2a/402a35a0-1222-4dab-b090-5c81bbf34237/declaracion_de_santiago.pdf?utm_source=chatgpt.com. Acesso em: 6 ago. 2026.

156 *Ibid*.

157 *Ibid*.

158 *Ibid*.

capable of generating incentives, accountability, and review pathways. [159]

A second feature is the centrality of the public sector (and, in some cases, the justice system) as the initial locus of institutionalization. In several countries, instruments concerning procurement, transparency, data protection, and internal governance precede, and sometimes replace for years, the enactment of comprehensive legislation. This dynamic converges with the notion of systemic regulation: when technology diffuses transversally, the State cannot act solely *ex post* in response to specific use cases; it must build *ex ante* capacities, such as impact assessments, documentation requirements, and audit trails. [160]

The third feature, decisive for the region, concerns the role of data and sovereignty. The contemporary debate on "*digital sovereignty*" is not limited to the territorial protection of the State; it also encompasses data governance as a condition for collective self-determination and individual autonomy. [161] In data-driven environments, AI regulation tends to produce extraterritorial effects, for instance, when interoperability standards and cross-border data flows begin to shape markets and services. Hence the relevance of incorporating, into the Latin American debate, clauses and institutions aimed at data control, transparency, and contestation mechanisms, lest the accelerated adoption of AI deepen technological dependencies.

Finally, recent scholarship emphasizes that, beyond principles, operationalization is essential: responsibilities, technical standards, certification procedures, and corrective mechanisms must be clearly defined. [162] In this regard, different authors

---

[159] CAJUEIRO, Daniel O.; CELESTINO, Victor R. R. *A Comprehensive Review of Artificial Intelligence Regulation*: Weighing Ethical Principles and Innovation. Available at: <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4950727>. Access: 28 Jan. 2025, *passim.*

[160] ARBEL, Yonathan; MATTHEW, Tokson; LIN, Albert. Systemic Regulation of Artificial Intelligence. *Arizona State Law Journal*. Phoenix: Arizona State University. v. 56, n. 2, p. 545-620, Summer, 2024,*passim.*

[161] SCASSA, Teresa. Sovereignty and the Governance of Artificial Intelligence. *UCLA Law Review*. Los Angeles: University of California School of Law. v. 71, n. 1, p. 214-229, May, 2024,*passim.*

[162] WOLF, Frank. *A Blueprint for the Regulation of AI Technologies*. 2022; CAJUEIRO, Daniel O.; CELESTINO, Victor R. R. *A Comprehensive Review of Artificial Intelligence Regulation*: Weighing Ethical Principles and Innovation. Available at: <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4950727>. Access: 28 Jan. 2025,*passim.*

propose models of trustworthy or responsible AI, centered on transparency, explainability, safety, human oversight, and ethical safeguards. [163] These vectors provide useful criteria for assessing whether the region is moving from a "market of principles" toward normative arrangements capable of reducing informational and power asymmetries among developers, the State, and citizens.

CONCLUSION

The 2026 regulatory mapping suggests that Latin America is passing through a phase of gradual consolidation rather than one of settled regulation. Across the ten jurisdictions examined, AI governance is exercised predominantly through strategic plans, guidelines, and administrative acts, with a strong incidence in the public sector, while legislative initiatives advance at markedly different speeds toward general and binding regimes. The categorisation adopted here, general or sectoral regulation, Executive or Legislative rulemaking, and the legal nature of the instrument, helps locate where institutional choices are effectively concentrated: more often in coordination bodies, programmes, and adoption protocols than in broad and stable statutory frameworks.

The comparison makes clear that no single regional model exists. What the survey reveals instead is a spectrum defined by the stage of regulatory maturation. At one end stands Peru, the only country with a general AI law already in force and regulated, complemented by a criminal-law reform and by a national strategy that itself concedes that the mere existence of rules does not guarantee coherent adoption by the state. At the other lies Argentina, which has no active bill and where the Executive has embraced an expressly deregulatory posture, leaving governance to soft-law instruments and provincial judicial protocols. Between these poles sit the majority: Brazil, Mexico, Colombia, Ecuador, Costa Rica, and Panama, each with a comprehensive bill under discussion but still resting, in the meantime, on strategies, sectoral reforms, or administrative acts.

---

[163] FLORIDI, Luciano; COWLS, Josh. *A unified framework of five principles for AI in society. Harvard Data Science Review*, Cambridge, MA, v. 1, n. 1, jul. 2019. Disponível em: https://hdsr.mitpress.mit.edu/pub/l0jsh9d1/release/8. Acesso em: 02 mar. 2025, *passim*.

Beneath this diversity, several convergences emerge. The first is the dominance of the risk-based approach, adopted in some form by every country that has drafted a general instrument, even as the number of tiers varies, two in Peru, Brazil, and Mexico, four in Chile, Colombia, Costa Rica, and Ecuador, and even as Costa Rica openly models its scheme on the European Union's AI Act while adjusting it to national circumstances. The second convergence concerns the dedicated AI authority: nearly every general bill provides for one, yet in most cases that authority exists only on paper, awaiting the enactment of the instrument that would create it. Only where governance already runs through an established agency, Peru's SGTD, Brazil's ANPD, Uruguay's AGESIC, does an operative body actually oversee the field today.

A third and more structural finding concerns the public sector and the judiciary as the initial locus of institutionalisation. In several countries, rules on transparency, data protection, and internal governance precede, and for years substitute for, comprehensive legislation. Brazil was the earliest to govern AI within its courts, through CNJ Resolution 332/2020, a comparatively spare instrument later deepened by the far more detailed Resolution 615/2025, while Colombia, with its Superior Council of the Judiciary guidelines of 2024, became the first country to implement the UNESCO recommendation on AI ethics within its judiciary. Peru reached a general statute earlier than any of them enacted one: its 2023 AI law was already in force while Brazil and Colombia were still building AI governance through judicial and administrative instruments in the absence of a comprehensive law. This dynamic converges with the notion of systemic regulation, in which the state, faced with a transversally diffusing technology, must build ex ante capacities rather than react case by case.

The fourth axis, decisive for the region, concerns data and sovereignty. The comparison shows two distinct ways of relating national regulation to the wider international order. Mexico affirms technological sovereignty as a limit on external alignment, subordinating its harmonisation with UNESCO, OECD, and USMCA instruments to national security and the domestic public interest; Brazil couples regulation with an infrastructural conception of sovereignty, extending to a sovereign cloud, national datasets, and Portuguese-language models. Uruguay moves in the opposite direction, becoming the first Latin American state to accede to the binding Council of Europe Convention on AI. That the same value, sovereignty, should yield both a reservation against international commitment and an early embrace of it captures,

better than any single instrument, the region's present heterogeneity.

A noteworthy point concerns neurorights, whose protection differs in rank between Chile and Mexico. Chile provides for it at the constitutional level, while Mexico approaches it through ordinary legislation, in Article 18 of its draft General AI Law, which recognises five such rights, a difference in hierarchy that also reflects a difference in stage, since the Mexican provision remains a bill.

The observed landscape reveals both promise and limitation. The predominance of flexible instruments fosters experimentation and technological adaptation, yet risks low normative density where it is not matched by clear accountability and enforcement mechanisms; a strategy or guideline without a legal basis may harden into a merely performative statement. The specialised literature indicates that consolidation requires the articulation of risk-proportionate ex ante instruments, robust data-protection safeguards, clarity as to responsibilities and review pathways, and some measure of regional coordination capable of mitigating fragmentation and technological dependency.

None of this points to a uniform Latin American model, and the criminal-law responses illustrate the point in miniature: where Peru treats the use of AI as a general aggravating factor across existing offences, Mexico proposes a dedicated offence of violation of sexual intimacy, and Brazil amends its Criminal Code to raise penalties for AI-altered image or voice, three techniques for a single problem. What emerges from the survey is a regulatory process still under construction, marked by institutional asymmetries and differing speeds of normative maturation. The regional challenge lies less in the formal adoption of principles. already a shared grammar across the ten countries, than in their translation into institutional arrangements able to balance technological innovation with the protection of fundamental rights.

REFERENCES

ALAGOAS. *Lei nº 9.095, de 11 de dezembro de 2023*: estabelece os princípios e diretrizes para o uso da inteligência artificial, no âmbito da administração pública estadual, e dá outras providências. Maceió: Assembleia Legislativa do Estado de Alagoas, 2023. Disponível em: https://sapl.al.al.leg.br/media/sapl/public/normajuridica/2023/2799/lei_no_9.095_de_11_de_dezembro_de_2023_.pdf. Acesso em: 06 ago. 2026.

ARBEL, Yonathan; TOKSON, Matthew; LIN, Albert. Systemic Regulation of Artificial Intelligence.

Arizona State Law Journal, Phoenix, v. 56, n. 2, p. 545-620, Summer 2024.

ARGENTINA. *Ley nº 25.326, de 4 de octubre de 2000*: Protección de los Datos Personales. Buenos Aires: Congreso de la Nación Argentina, 2000. Disponível em: https://servicios.infoleg.gob.ar/infolegInternet/anexos/60000-64999/64790/texact.htm. Acesso em: 06 ago. 2026.

ARGENTINA. Poder Ejecutivo Nacional. *Mensaje: Ley General de Sociedades*. Buenos Aires, 2026. Projeto de lei encaminhado ao Honorable Congreso de la Nación. Disponível em: https://aldiaargentina.microjuris.com/wp-content/uploads/2026/06/proyecto-ley-general-sociedades.pdf?ref=digital-nomos.ghost.io. Colegio Público de la Abogacía de la Capital Federal – documento em PDF. Acesso em: 6 ago. 2026.

ARGENTINA. Presidencia de la Nación. *Plan Nacional de Inteligencia Artificial*. Buenos Aires: Presidencia de la Nación, 2019. Disponível em: https://api.oecdai.org/storage//policy-initiatives/Jul2025/fu_rh7qgrkz9xkx89g.pdf. Acesso em: 6 ago. 2026.

BALDWIN, Robert; CAVE, Martin; LODGE, Martin. *Understanding Regulation*: Theory, Strategy, and Practice. Oxford: Oxford University Press, 2012.

BRASIL. Conselho Nacional de Justiça. *Resolução nº 615, de 11 de março de 2025*: estabelece diretrizes para o desenvolvimento, utilização e governança de soluções desenvolvidas com recursos de inteligência artificial no Poder Judiciário. Brasília, DF: CNJ, 2025. Disponível em: https://atos.cnj.jus.br/files/original1555302025031467d4517244566.pdf. Acesso em: 06 ago. 2026.

BRASIL. *Lei nº 13.709, de 14 de agosto de 2018*: Lei Geral de Proteção de Dados Pessoais (LGPD). Brasília, DF: Presidência da República, 2018. Disponível em: https://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/l13709.htm. Acesso em: 02 ago. 2026.

BRASIL. *Lei nº 14.129, de 29 de março de 2021*: dispõe sobre princípios, regras e instrumentos para o Governo Digital e para o aumento da eficiência pública. Brasília, DF: Presidência da República, 2021. Disponível em: https://www.planalto.gov.br/ccivil_03/_ato2019-2022/2021/lei/l14129.htm. Acesso em: 02 ago. 2026.

BRASIL. *Lei nº 15.123, de 24 de abril de 2025*: altera o art. 147-B do Decreto-Lei nº 2.848, de 7 de dezembro de 1940 (Código Penal), para estabelecer causa de aumento de pena no crime de violência psicológica contra a mulher quando praticado com o uso de inteligência artificial ou de qualquer outro recurso tecnológico que altere imagem ou som da vítima. Brasília, DF: Presidência da República, 2025. Disponível em: https://www.planalto.gov.br/ccivil_03/_ato2023-2026/2025/lei/l15123.htm. Acesso em: 06 ago. 2026.

BRASIL. *Lei nº 15.211, de 17 de setembro de 2025*: dispõe sobre a proteção de crianças e adolescentes em ambientes digitais (ECA Digital). Brasília, DF: Presidência da República, 2025. Disponível em: https://www.planalto.gov.br/ccivil_03/_ato2023-2026/2025/lei/l15211.htm. Acesso em: 02 ago. 2026.

BRASIL. Ministério da Ciência, Tecnologia e Inovação. *Plano Brasileiro de Inteligência Artificial*

(PBIA): IA para o bem de todos. Brasília, DF: MCTI, 2024. Disponível em: https://www.gov.br/mcti/pt-br/acompanhe-o-mcti/transformacaodigital/plano-brasileiro-de-inteligencia-artificial-pbia-_vf.pdf. Acesso em: 02 ago. 2026.

BRASIL. Senado Federal. *Projeto de Lei nº 2.338, de 2023*: dispõe sobre o uso da inteligência artificial. Brasília, DF: Senado Federal, 2023. Disponível em: https://legis.senado.leg.br/sdleg-getter/documento?dm=9915448&ts=1774551291270&disposition=inline. Acesso em: 02 ago. 2026.

BRASIL. Tribunal Superior Eleitoral. *Resolução nº 23.670, de 14 de dezembro de 2021*: dispõe sobre as federações de partidos políticos. Brasília, DF: TSE, 2021. Disponível em: https://www.tse.jus.br/legislacao/compilada/res/2021/resolucao-no-23-670-de-14-de-dezembro-de-2021. Acesso em: 06 ago. 2026.

CADENA POLÍTICA. *México: reforma a la Ley Federal del Trabajo y a la Ley Federal del Derecho de Autor sobre inteligencia artificial y artistas*. Cadena Política, 15 maio 2026. Disponível em: https://cadenapolitica.com/2026/05/15/mexico-reforma-ley-federal-trabajo-derecho-autor-ia-artistas. Acesso em: 06 ago. 2026.

CAJUEIRO, Daniel O.; CELESTINO, Victor R. R. *A Comprehensive Review of Artificial Intelligence Regulation*: Weighing Ethical Principles and Innovation. Disponível em: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4950727. Acesso em: 28 jan. 2025.

CHILE. *Ley 21.663, Ley Marco de Ciberseguridad*. Santiago: Biblioteca del Congreso Nacional de Chile, 2024. Disponível em: http://bcn.cl/leychile/navegar?idNorma=1202434. Acesso em: 6 ago. 2026.

CHILE. *Ley 21.719, que regula la protección y el tratamiento de los datos personales y crea la Agencia de Protección de Datos Personales*. Santiago: Biblioteca del Congreso Nacional de Chile, 2024. Disponível em: https://www.bcn.cl/leychile/navegar?idNorma=1209272. Acesso em: 6 ago. 2026.

CHILE. Cámara de Diputadas y Diputados. *Proyecto de ley que regula los sistemas de inteligencia artificial*: boletín n. 16.821-19. Santiago: Cámara de Diputadas y Diputados. Disponível em: https://www.camara.cl/legislacion/proyectosdeley/tramitacion.aspx?prmID=17429&prmBOLETIN=16821-19. Acesso em: 6 ago. 2026.

COLÔMBIA. Departamento Nacional de Planeación. *Documento CONPES 4144*: Política Nacional de Inteligencia Artificial. Bogotá: DNP, 2025. Disponível em: https://siteal.iiep.unesco.org/sites/default/files/sit_accion_files/documento_conpes_4144_-_politica_nacional_de_inteligencia_artificial.pdf. Acesso em: 06 ago. 2026.

COLÔMBIA. *Estrategia Nacional Digital de Colombia 2023–2026*. Bogotá: Colombia Government, 2023. Disponível em: https://siteal.iiep.unesco.org/sites/default/files/sit_accion_files/estrategia_nacional_digital_de_colombia_2023-2026.pdf. Acesso em: 06 ago. 2026.

COLÔMBIA. *Ley 1581 de 2012*: por la cual se dictan disposiciones generales para la protección de datos personales. Bogotá: Congreso de la República, 2012. Disponível em: https://www.funcionpublica.gov.co/eva/gestornormativo/norma.php?i=49981. Acesso em: 06 ago. 2026.

COLÔMBIA. Ministerio de Ciencia, Tecnología e Innovación. *Hoja de ruta para el desarrollo y aplicación de la inteligencia artificial en Colombia*. Bogotá: MinCiencias, 2024. Disponível em: https://minciencias.gov.co/sites/default/files/upload/noticias/hoja_de_ruta_adopcion_etica_y_sostenible_de_inteligencia_artificial_colombia_0.pdf. Acesso em: 06 ago. 2026.

COLÔMBIA. Ministerio de Ciencia, Tecnología e Innovación. *Proyecto de Ley nº 422 de 2025*. Bogotá: MinCiencias, 2025. Disponível em: https://minciencias.gov.co/sites/default/files/upload/noticias/pl_ia_finalizado.pdf. Acesso em: 06 ago. 2026.

COSTA RICA. Asamblea Legislativa. *Proyecto de Ley nº 23.771*. San José: Asamblea Legislativa de la República de Costa Rica, 2023. Disponível em: https://delfino.cr/asamblea/proyecto/23771. Acesso em: 06 ago. 2026.

COSTA RICA. Ministerio de Ciencia, Innovación, Tecnología y Telecomunicaciones. *National Artificial Intelligence Strategy of Costa Rica*. San José: MICITT, 2024. Disponível em: https://www.micitt.go.cr/sites/default/files/2025-02/National%20Artificial%20Intelligence%20Strategy%20of%20Costa%20Rica.pdf. Acesso em: 06 ago. 2026.

CÚPULA MINISTERIAL E DE ALTAS AUTORIDADES DA AMÉRICA LATINA E DO CARIBE. *Declaración de Santiago*: para promover una inteligencia artificial ética en América Latina y el Caribe. Santiago de Chile, 23-24 out. 2023. Disponível em: https://minciencia.gob.cl/uploads/filer_public/40/2a/402a35a0-1222-4dab-b090-5c81bbf34237/declaracion_de_santiago.pdf. Acesso em: 6 ago. 2026.

DIPUTADOS de Morena anuncian iniciativa para facultar al Congreso legislar sobre la IA. *La Jornada*, Ciudad de México, 6 ago. 2026. Política. Disponível em: https://www.jornada.com.mx/noticia/2026/08/08/politica/diputados-de-morena-anuncia-iniciativa-para-facultar-al-congreso-legislar-sobre-la-ia. Acesso em: 06 ago. 2026.

EQUADOR. Asamblea Nacional. *Proyecto de Ley Orgánica de Regulación y Promoción de la Inteligencia Artificial en Ecuador*. Quito, 2024. Disponível em: https://www.asambleanacional.gob.ec/sites/default/files/private/asambleanacional/filesasambleanacionalnameuid-19130/2192.%20Proyecto%20de%20Ley%20Organica%20de%20Regulacion%20y%20Promocion%20de%20la%20Inteligencia%20Artificial%20en%20Ecuador%20-pnunez/pp%20-%20proyecto%20de%20ley%20450889-nunez.pdf. Acesso em: 06 ago. 2026.

EQUADOR. Superintendencia de Competencia Económica. *Decálogo ético para el uso de la inteligencia artificial*. 2026. Disponível em: https://oecd.ai/en/dashboards/policy-initiatives/ethical-decalogue-for-the-use-of-artificial-intelligence-ai. Acesso em: 06 ago. 2026.

FLORIDI, Luciano; COWLS, Josh. A unified framework of five principles for AI in society. *Harvard Data Science Review*, Cambridge, MA, v. 1, n. 1, jul. 2019. Disponível em: https://hdsr.mitpress.mit.edu/pub/l0jsh9d1/release/8. Acesso em: 02 mar. 2025.

GOIÁS. *Lei Complementar nº 205, de 19 de maio de 2025*: institui a Política Estadual de Fomento à Inovação em Inteligência Artificial no Estado de Goiás. Goiânia: Casa Civil do Estado de Goiás,

2025. Disponível em: https://legisla.casacivil.go.gov.br/api/v2/pesquisa/legislacoes/110694/pdf. Acesso em: 06 ago. 2026.

GUIHOT, Michael; MATTHEW, Anne F.; SUZOR, Nicolas. *Nudging Robots*: Innovative Solutions to Regulate Artificial Intelligence. Vanderbilt Journal of Entertainment & Technology Law, no prelo, 28 jul. 2017. Disponível em: https://ssrn.com/abstract=3017004. Acesso em: 02 mar. 2025.

HARASIMIUK, Dominika Ewa; BRAUN, Tomasz. *Regulating artificial intelligence*: binary ethics and the law. Abingdon: Routledge, 2021.

INTRONA, Lucas D. Algorithms, governance, and governmentality: on governing academic writing. *Science, Technology, & Human Values*, v. 41, n. 1, p. 17-49, 2016. Disponível em: http://www.jstor.org/stable/43671281. Acesso em: 02 mar. 2025.

JUDGE, Brian; NITZBERG, Mark; RUSSELL, Stuart. When code isn't law: rethinking regulation for artificial intelligence. *Policy and Society*, Oxford, p. 01-13, maio 2024.

KARTSKHIYA, Aleksandr Amiranovich. Legal horizons of the new artificial intelligence paradigm. *Legal Issues in the Digital Age*, Moscow, v. 5, n. 2, p. 4-35, jul. 2024.

LÓPEZ MARTÍNEZ, Víctor José. Hacia la ley mexicana de la inteligencia artificial y sus derivados: buques autónomos, drones y robots. *El Financiero*, Ciudad de México, 3 ago. 2026. Opinión. Disponível em: https://www.elfinanciero.com.mx/opinion/victor-jose-lopez-martinez/2026/08/03/hacia-la-ley-mexicana-de-la-inteligencia-artificial-y-sus-derivados-buques-autonomos-drones-y-robots/. Acesso em: 06 ago. 2026.

MAGRANI, Eduardo. Novas perspectivas sobre ética e as leis da inteligência artificial. *Internet Policy Review*, v. 8, n. 3, p. 1-19, set. 2019. DOI: 10.14763/2019.3.1420.

MÉXICO. *Constitución Política de los Estados Unidos Mexicanos*. Ciudad de México: Suprema Corte de Justicia de la Nación, [2026]. Disponível em: https://www.scjn.gob.mx/sites/default/files/cpeum/documento/cpeum.pdf. Acesso em: 06 ago. 2026.

MÉXICO. *Ley Federal de Protección de Datos Personales en Posesión de los Particulares*. Nueva ley publicada en el Diario Oficial de la Federación el 20 de marzo de 2025. Última reforma publicada DOF 14-11-2025. Disponível em: https://dgesui.ses.sep.gob.mx/sites/default/files/2026-01/LFPDPPP.pdf. Acesso em: 06 ago. 2026.

MÉXICO. *Ley Federal del Derecho de Autor, de 24 de diciembre de 1996*. Última reforma publicada DOF 14-05-2026. Ciudad de México: Cámara de Diputados, 2026. Disponível em: https://www.diputados.gob.mx/LeyesBiblio/pdf/LFDA.pdf. Acesso em: 06 ago. 2026.

MÉXICO. *Ley Federal del Trabajo, de 1º de abril de 1970*. Última reforma publicada DOF 14-05-2026. Ciudad de México: Cámara de Diputados, 2026. Disponível em: https://www.diputados.gob.mx/LeyesBiblio/pdf/LFT.pdf. Acesso em: 06 ago. 2026.

MÉXICO. Secretaría de Gobernación. Sistema de Información Legislativa. *Iniciativa que reforma el artículo 199 Octies del Código Penal Federal, en materia de violencia digital, protección a la intimi-*

dad y sanción del uso indebido de tecnologías e inteligencia artificial: asunto n. 5.015.968. Ciudad de México, 17 fev. 2026. Disponível em: https://sil.gobernacion.gob.mx/Archivos/Documentos/2026/02/asun_5015968_20260217_1771373777.pdf. Acesso em: 06 ago. 2026.

MÉXICO. Senado de la República. Comisión de Análisis, Seguimiento y Evaluación sobre la Aplicación y Desarrollo de la Inteligencia Artificial en México. *Iniciativa con proyecto de decreto por el que se expide la Ley General para Regular y Fomentar el Uso de la Inteligencia Artificial (IA) en México*. Ciudad de México: LXVI Legislatura del Congreso de la Unión, mar. 2026.

MILEI defende IA desregulada após alerta de historiador Harari. *UOL Notícias*, São Paulo, 8 jun. 2026. Disponível em: https://noticias.uol.com.br/ultimas-noticias/afp/2026/06/08/milei-defende-ia-desregulada-apos-alerta-de-historiador-harari.htm. Acesso em: 06 ago. 2026.

OECD.AI. *National AI policies: Chile*. Paris: OECD, 2026. Disponível em: https://oecd.ai/en/dashboards/national/chile. Acesso em: 6 ago. 2026

OXFORD INSTITUTE OF TECHNOLOGY AND JUSTICE. *Argentina*. Oxford: University of Oxford, 2026. Disponível em: https://www.techandjustice.bsg.ox.ac.uk/research/argentina. Acesso em: 6 ago. 2026.

PANAMÁ. *AI 2036: National Artificial Intelligence Strategy of Panama*. Ciudad de Panamá, 2026. Disponível em: https://aipanama.gob.pa/. Acesso em: 06 ago. 2026.

PANAMÁ. Asamblea Nacional. *Panamá a la vanguardia de la tecnología con IA, según proyecto de ley aprobado en primer debate*. Ciudad de Panamá, 2026. Disponível em: https://www.asamblea.gob.pa/Noticias/Noticias/PANAMA-A-LA-VANGUARDIA-DE-LA-TECNOLOGIA-CON-IA-SEGUN-PROYECTO-DE-LEY-APROBADO-EN-PRIMER-DEBATE. Acesso em: 06 ago. 2026.

PANAMÁ. Secretaría Nacional de Ciencia, Tecnología e Innovación. *Panamá lanza la Agenda Nacional de Tecnologías Avanzadas en presencia del presidente Mulino*. Ciudad de Panamá, 30 jul. 2026. Disponível em: https://www.senacyt.gob.pa/en/panama-lanza-la-agenda-nacional-de-tecnologias-avanzadas-en-presencia-del-presidente-mulino/. Acesso em: 06 ago. 2026.

PANAMÁ. Secretaría Nacional de Ciencia, Tecnología e Innovación. *Plan Estratégico Nacional de Ciencia, Tecnología e Innovación (PENCYT) 2025-2029*. Ciudad de Panamá: SENACYT, 2025. Disponível em: https://www.senacyt.gob.pa/wp-content/uploads/2025/05/PENCYT-2025-2029.pdf. Acesso em: 06 ago. 2026.

PARANÁ. *Lei nº 22.343, de 04 de abril de 2025*: dispõe sobre a regulamentação do uso da Inteligência Artificial no Estado do Paraná. Curitiba: Assembleia Legislativa do Estado do Paraná, 2025. Disponível em: https://www.legisweb.com.br/legislacao/?id=476131. Acesso em: 06 ago. 2026.

PERU. *Decreto Legislativo n. 1412, que aprueba la Ley de Gobierno Digital*. Diario Oficial El Peruano, Lima, 2018. Disponível em: https://cdn.www.gob.pe/uploads/document/file/353216/decreto-legislativo-que-aprueba-la-ley-de-gobierno-digital-decreto-legislativo-n-1412-1691026-1.pdf. Acesso em: 06 ago. 2026.

PERU. *Decreto Supremo n. 115-2025-PCM, que aprueba el Reglamento de la Ley n. 31814*, Ley que promueve el uso de la inteligencia artificial en favor del desarrollo económico y social del país. Diario Oficial El Peruano, Lima, 2025. Disponível em: https://busquedas.elperuano.pe/dispositivo/NL/2436426-1. Acesso em: 06 ago. 2026.

PERU. *Estrategia Nacional de Inteligencia Artificial 2026-2030*. Lima: PCM, 2026. Disponível em: https://cdn.www.gob.pe/uploads/document/file/10347430/8408055-estrategia-nacional-de-inteligencia-artificial-2026-2030.pdf. Acesso em: 06 ago. 2026.

PERU. *Ley n. 31814, que promueve el uso de la inteligencia artificial en favor del desarrollo económico y social del país*. Lima, 2023. Disponível em: https://cdn.www.gob.pe/uploads/document/file/5038703/ley-que-promueve-el-uso-de-la-inteligencia-artificial-en-fav-ley-n-31814.pdf. Acesso em: 06 ago. 2026.

PERU. *Ley n. 32314, que modifica el Código Penal y la Ley de Delitos Informáticos para incorporar el uso de inteligencia artificial como agravante en la comisión de diversos delitos*. Diario Oficial El Peruano, Lima, 2025. Disponível em: https://busquedas.elperuano.pe/dispositivo/NL/2394851-2. Acesso em: 06 ago. 2026.

PERU. Presidencia del Consejo de Ministros. *Política Nacional de Transformación Digital al 2030*. Lima: PCM, 2023. Disponível em: https://cdn.www.gob.pe/uploads/document/file/4912655/Politica%20Nacional%20de%20Transformacion%20Digital%20al%202030_Resumen%20ejecutivo.pdf. Acesso em: 06 ago. 2026.

RIO DE JANEIRO (Município). *Lei nº 8.926, de 2025*: institui o Programa sobre o Uso Responsável da Inteligência Artificial no âmbito do Município e dá outras providências. Rio de Janeiro: Câmara Municipal do Rio de Janeiro, 2025. Disponível em: https://leis.org/municipais/rj/rio-de-janeiro/lei/lei-ordinaria/2025/8926/. Acesso em: 06 ago. 2026.

SAXENA, Ashish K. AI in Governance and Policy Making. *International Journal of Science and Research (IJSR)*, v. 13, n. 5, p. 1218-1223, maio 2024. DOI: 10.21275/SR24519015426. Disponível em: https://www.ijsr.net/getabstract.php?paperid=SR24519015426. Acesso em: 02 mar. 2025.

SCASSA, Teresa. Sovereignty and the Governance of Artificial Intelligence. *UCLA Law Review*, Los Angeles, v. 71, n. 1, p. 214-229, May 2024.

TECH AND JUSTICE. *Colombia*. Oxford: Blavatnik School of Government, University of Oxford, 2026. Disponível em: https://www.techandjustice.bsg.ox.ac.uk/research/colombia. Acesso em: 06 ago. 2026.

UNESCO. *Chile: pioneering the protection of neurorights*. The UNESCO Courier, Paris, 2022. Disponível em: https://courier.unesco.org/en/articles/chile-pioneering-protection-neurorights. Acesso em: 6 ago. 2026.

UNESCO. *Ethics of artificial intelligence: Mexico*. Paris: UNESCO, 2026. Disponível em: https://www.unesco.org/ethics-ai/en/mexico. Acesso em: 06 ago. 2026.

UNESCO. *Ethics of artificial intelligence: Panama*. Paris: UNESCO, 2026. Disponível em: https://www.unesco.org/ethics-ai/en/panama. Acesso em: 06 ago. 2026.

UNESCO. *Ethics of artificial intelligence: Uruguay*. Paris: UNESCO, 2026. Global AI Ethics and Governance Observatory. Disponível em: https://www.unesco.org/ethics-ai/en/uruguay. Acesso em: 06 ago. 2026.

UNESCO. *Recomendação sobre a Ética da Inteligência Artificial*. Paris: UNESCO, 2022. Disponível em: https://unesdoc.unesco.org/ark:/48223/pf0000381137_por. Acesso em: 02 mar. 2025.

UNIÃO EUROPEIA. Comissão Europeia. *Adequacy decisions*. Brussels: European Commission, [s. d.]. Disponível em: https://commission.europa.eu/law/law-topic/data-protection/international-dimension-data-protection/adequacy-decisions_en. Acesso em: 06 ago. 2026.

URUGUAI. Agencia de Gobierno Electrónico y Sociedad de la Información y del Conocimiento. *Aprobación de la Declaración de Montevideo*. Montevidéu: Agesic, 2024. Disponível em: https://www.gub.uy/agencia-gobierno-electronico-sociedad-informacion-conocimiento/comunicacion/noticias/aprobacion-declaracion-montevideo. Acesso em: 06 ago. 2026.

URUGUAI. Agencia de Gobierno Electrónico y Sociedad de la Información y del Conocimiento. *Estrategia Nacional de Inteligencia Artificial del Uruguay 2024-2030*. Montevidéu: Agesic, 2024. Disponível em: https://www.gub.uy/agencia-gobierno-electronico-sociedad-informacion-conocimiento/comunicacion/publicaciones/estrategia-nacional-inteligencia-artificial-del-uruguay-2024-2030. Acesso em: 06 ago. 2026.

URUGUAI. Agencia de Gobierno Electrónico y Sociedad de la Información y del Conocimiento. *Tratado del Consejo de Europa sobre IA: Uruguay es el primer país de Latinoamérica en adherirse*. Montevidéu: Agesic, set. 2025. Disponível em: https://www.gub.uy/agencia-gobierno-electronico-sociedad-informacion-conocimiento/comunicacion/noticias/tratado-del-consejo-europa-sobre-ia-uruguay-es-primer-pais-latinoamerica. Acesso em: 06 ago. 2026.

URUGUAI. *Ley 20.212, de 6 de noviembre de 2023*: rendición de cuentas y balance de ejecución presupuestal, ejercicio 2022. Artículo 74. Montevidéu: IMPO, 2023. Disponível em: https://www.impo.com.uy/bases/leyes/20212-2023/74. Acesso em: 06 ago. 2026.

WALTERS, William. Some critical notes on "governance". *Studies in Political Economy*, v. 73, 2004. Disponível em: https://www.tandfonline.com/doi/abs/10.1080/19187033.2004.11675150. Acesso em: 04 mar. 2025.

WOLF, Frank. *A blueprint for the regulation of artificial intelligence technologies*. Bradford: Ethics International Press, 2022.

THE NEW LEGISLATIVE FRAMEWORK (NLF) AND THE DANGER OF DECONTEXTUALIZED IMPORTATION OF THE ARTIFICIAL INTELLIGENCE ACT TO BRAZIL

Heitor Pedrosa Haddad

Master's Student in Law (UFMG)

DOI: https://doi.org/10.59224/dti8.ch9

Abstract: The article analyzes Regulation (EU) 2024/1689 (AI Act) in light of the New Legislative Framework (NLF), demonstrating its integration into a regulatory model focused on product safety and the consolidation of the European internal market, a framework known as the New Approach. It contrasts with the GDPR, which prioritizes fundamental rights. The study also examines the influence of the AI Act on Brazil's Bill No. 2,338/2023, warning of the risks of its decontextualized "importation". Accordingly, the study seeks to examine the specificities of the Brazilian regulatory framework concerning product safety. A dogmatic-legal approach is adopted, employing the hypothetical-deductive method and relying on normative and doctrinal sources.

Keywords: New Legislative Framework - Artificial Intelligence Act – Product Safety Regulation

Table of contents: 1. Introduction. 2. Methodology. 3. Old Approach x New Approach. 4. New Legislative Framework (NLF). 4.1. Regulation 765/2008, Regulation 2019/1020 and Decision nº 768/2008/CE. 4.2. Evaluation of the New Legislative Framework. 5. Artificial Intelligence Act x General Data Protection Regulation. 5.1. Artificial Intelligence Act. 5.2. General Data Protection Regulation. 5.3. Conclusion. 6. Brazilian Bill 2,338/23. 7. The Product Safety System in Brazil. 8. Conclusion.

1. INTRODUCTION

This article aims to investigate the legal-normative context in which Regulation (EU) 2024/1689 (Artificial Intelligence Act – AI Act) is situated. It starts from the hypothesis that many of its provisions and regulatory modeling are closely linked to the structural logic of the New Legislative Framework (NLF), which allows understanding the AI Act not as an isolated norm, but as part of a broader legal system aimed at normative harmonization among European Union Member States in the field of product safety and the consolidation of the internal market.

The work initially proposes an analysis of the NLF, understood as the unfolding of the paradigm inaugurated by the New Approach, investigated in Chapter 3. The main provisions of Regulation (EC) No. 765/2008, Regulation (EU) 2019/1020, and Decision No. 768/2008/EC, central acts of the NLF, will be examined in Chapter 4. The investigation will be based both on a direct reading of the legal texts and on an analysis of the Blue Guide, an interpretive document prepared by the European Commission, as well as on evaluative reports produced by the same institution after years of the regulatory framework's enforcement, which reflect the perception of civil society, economic operators, and public authorities regarding the effectiveness of the NLF.

In Chapter 5, the article contextualizes the AI Act within the NLF, highlighting how several of its provisions reflect the regulatory structure established by the mentioned legal instruments. In parallel, a comparison is made with the General Data Protection Regulation (GDPR), a norm that, not being part of the NLF, is structured differently, focusing primarily on the protection of fundamental rights.

The analysis will also include a discussion about the so-called "Brussels Effect," especially in the context of the GDPR, considered one of the European Union's most influential legal instruments in terms of global normative projection. In Chapter 6, the presence of traits of the AI Act in Bill No. 2,338/2023, currently under consideration in the Brazilian National Congress, is identified, although this bill moves towards a more fundamental rights-centered approach.

The goal is to demonstrate that, regardless of the validity of the risk-based approach as a regulatory method, it is essential that the Brazilian legislator understands

the specificities of the legal framework in which the AI Act is inserted. Not only that, it is fundamental to understand Brazil's product safety regulation system, as pointed out in Chapter 7. Otherwise, there is a serious risk of decontextualized transposition of some parts of the AI Act into the Brazilian legal system.

2. METHODOLOGY

The research adopts a dogmatic-legal approach, considering law with methodological self-sufficiency. The reasoning method employed was hypothetical-deductive, with the formulation of an initial hypothesis to be refined in light of normative and doctrinal analysis. It is, therefore, theoretical research, of the legal-comprehensive type, whose objective is to decompose, understand, and critically interpret the legal-normative context in which the AI Act is inserted.

The data used were predominantly primary, given that the analysis is largely based on normative acts. Complementarily, secondary sources, especially specialized literature, will be used to support normative interpretation and provide theoretical backing for the debate. It is noteworthy that the examination of European legislations was conducted, throughout the study, with the interpretive support offered by the *Blue Guide on the Implementation of EU Product Rules*. This approach allowed a more comprehensive understanding of the examined norms, inserting them into their practical application context and highlighting their systematic connections with other related normative instruments.

The data collection and analysis procedure consisted of a bibliographic and documentary survey, with emphasis on works discussed in the discipline "Artificial Intelligence Regulation Landscape" taught by Professor Leonardo Parentoni, as well as consultation with current legislation and relevant legislative proposals on the topic.

3. OLD APPROACH X NEW APPROACH

To understand the *raison d'être* of the Artificial Intelligence Act, it is essential to place it within the broader normative context of the European Union, marked by the progressive consolidation, integration, and development of its internal market. Such understanding requires an analysis of the evolution of the European regulatory

framework, especially the paradigmatic transition from the so-called Old Approach, centered on national legislations and aimed at protecting each country's internal markets, to the New Approach, which introduced a new regulatory logic oriented towards technical harmonization and the free movement of goods within the European space. This process of normative transformation is, in turn, deepened and systematized through the New Legislative Framework (NLF), which strengthens the European Union's legal and institutional instruments with a view to consolidating a functional single market, in which the AI Act finds its foundation.

Traditionally, in the Old Approach model, the technical requirements for the international commercialization of products were dictated by national legislation, with often very detailed rules, which was responsible for the emergence of legislative divergences among European Union countries; consequently, trade between Member States was hampered[1].

Therefore, subsequently, the focus of the New Approach was to ensure that Member States did not obstruct the free movement of goods by imposing national technical standards, encouraging regulatory compatibility. To this end, the following principles were established: legislative harmonization should only focus on essential technical requirements for products, so as to remove unjustified restrictions on the movement of goods; technical specifications of products that meet the essential requirements established in the legislation must be defined in harmonized standards; products manufactured in conformity with harmonized standards enjoy a presumption of conformity with the corresponding essential requirements; the application of harmonized standards or other technical standards remains voluntary, and the manufacturer can always use other technical specifications to meet the requirements, but, in this case, will assume the burden of demonstrating that these technical specifications effectively meet the essential requirements set out in law, which often requires the intervention of a conformity assessment body[2].

---

1 DU BOISPÉAN, Stéphane; MUECK, Markus; GAIE, Christoph. *Introduction to the European New Legislative Framework*. In: MUECK, Markus; GAIE, Christoph (ed.). European Digital Regulations. Cham: Springer, 2025. (Intelligent Systems Reference Library, v. 265). Available at: https://doi.org/10.1007/978-3-031-80809-8_1.

2 EUROPEAN COMMISSION. *Blue Guide on the implementation of EU product rules 2022*. Luxembourg: Publications Office of the European Union, 2022. Available at: https://eur-lex.euro-

In this context, only the essential requirements aimed at the safety of a given product are normatively defined by the European Union. The results to be achieved or the risks to be balanced are determined, but the technical solutions aimed at concretizing such objectives are not specified[3]. These, in turn, are left to harmonized standards[4], which are optional, but capable of generating a presumption of conformity with an essential requirement.

These harmonized European Union standards can be totally or partially based on international standards, such as those developed by ISO (International Organization for Standardization) or IEC (International Electrotechnical Commission). However, the presumption of conformity with the essential requirements established in European Union legislation is only granted when the European version of the recommendation is adopted, formally published in the *Official Journal of the European Union*. This is due to the fact that, on many occasions, the original versions of ISO or IEC standards do not establish a direct correlation with the essential requirements provided for in European legislation[5]. In contrast, harmonized standards are specifically developed for this purpose, to allow products or systems to presumably meet essential requirements, generally dispensing with additional conformity analyses.

If economic operators choose not to follow a harmonized standard or to apply it only partially, they bear the burden of demonstrating, by alternative means of their choice, that their products fully meet the essential requirements. This possibility, although legitimate, implies the assumption of greater technical and legal responsibility on the part of manufacturers, due to the absence of the presumption of conformity that harmonized standards provide[6].

---

pa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52022XC0629(04). Acess in: 16 June 2025.

3 EUROPEAN COMMISSION, Blue Guide, 2022.

4 A harmonized standard is a European standard developed by a recognized European standardization body: CEN, CENELEC, or ETSI. It is created following a request from the European Commission to one of these organizations. Manufacturers, other economic operators, or conformity assessment bodies can use harmonized standards to demonstrate that products, services, or processes comply with relevant EU legislation. (EUROPEAN COMMISSION. *Harmonised standards*)

5 EUROPEAN COMMISSION, *Blue Guide*, 2022.

6 EUROPEAN COMMISSION, *Blue Guide*, 2022.

By way of illustration, Directive 2009/48/EC, concerning the safety of toys, establishes essential safety requirements to be observed by manufacturers, which are introduced by Article 10 and detailed in Annex II. Article 13 of the same directive, under the heading "presumption of conformity," states that toys shall be presumed to be in conformity with the requirements of Article 10 and Annex II when manufactured in conformity with harmonized standards, the references of which have been published in the *Official Journal of the European Union*[7]. Subsequently, such harmonized standards were indeed published with the specific objective of allowing the presumption of conformity in relation to the provisions of the toy safety directive[8].

The Artificial Intelligence Act follows the same logic. Under Article 40, §1, high-risk artificial intelligence systems, as well as general-purpose AI models (GPAI), are considered to be in conformity with the requirements set out in Section 2 of Chapter III or, where applicable, with the obligations contained in Sections 2 and 3 of Chapter V, when they adhere to harmonized standards whose references have been published in the *Official Journal of the European Union*[9].

Entities such as the European Committee for Standardization (CEN), the European Electrotechnical Committee for Standardization (CENELEC), and the European Telecommunications Standards Institute (ETSI) are responsible for elaborating these complementary technical standards of a non-binding nature. By June 2025, fifteen

---

7 EUROPEAN UNION. Directive 2009/48/EC of the European Parliament and of the Council of 18 June 2009 on the safety of toys. *Official Journal of the European Union*, 30 June 2009. Available at: https://eur-lex.europa.eu/legal-content/PT/TXT/?uri=CELEX:32009L0048. Access in: 17 June 2025.

8 EUROPEAN UNION. European Commission. Commission Implementing Decision (EU) 2023/740 of 4 April 2023 on harmonised standards for toys drafted in support of Directive 2009/48/EC of the European Parliament and of the Council. *Official Journal of the European Union*, 5 April 2023.Available at: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32023D0740. Access in: 17 June 2025.)

9 EUROPEAN UNION. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending Regulations (EC) No 300/2008, (EU) No 2018/858, (EU) 2019/2144, (EU) 2019/881 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828. *Official Journal of the European Union*, 12 June 2024. Available at: https://eur-lex.europa.eu/legal-content/PT/TXT/?uri=CELEX%3A32024R1689. Access in: 17 June 2025.

harmonized standards aimed at the application of the AI Act had already been published by CENELEC[10].

4. NEW LEGISLATIVE FRAMEWORK (NLF)

The New Legislative Framework (NLF) represents a normative evolution of the New Approach, also constituting a structured initiative of the European Union aimed at harmonizing rules related to product safety. Its fundamental objective is to facilitate and intensify the commercialization of goods within the European market, promoting the elimination of technical barriers between Member States[11].

Its emergence dates back to 2008, consolidating the fundamental role of technical standards and patterns adopted or referenced at the European level as facilitators of economic integration. Thus, the competence to regulate the placing of products on the market materialized even more from member countries to the European level[12].

It is relevant to note that the process of intensifying the European Union's normative action, through the production of supranational norms, is intrinsically related to the evolution of the functions traditionally attributed to Directives and Regulations, two recurring legal instruments in the European scenario.

Directives, as a rule, are legal acts that need to be transposed into the internal legal order of Member States to produce full legal effects. Regulations, in turn, have direct and immediate applicability in all Member States, regardless of any measure of incorporation into national law. In recent decades, however, the increase in both the scope and normative precision of Directives leaves little room for countries to innovate

---

[10] CENELEC – European Committee for Electrotechnical Standardization. European Committee for Electrotechnical Standardization (CENELEC). Bruxelas: CENELEC, 2025. Available at: https://standards.cencenelec.eu/dyn/www/f?p=205:32:0::::FSP_ORG_ID,FSP_LANG_ID:2916257,25&cs=1827B89DA69577BF3631EE2B6070F207D. Access in: 17 June 2025.

[11] EUROPEAN COMMISSION. *New Legislative Framework*. Single Market – European Commission. Available at: https://single-market-economy.ec.europa.eu/single-market/goods/new-legislative-framework_en. Access in: 16 June 2025.

[12] DU BOISPÉAN; MUECK; GAIE, *Introduction to the European New Legislative Framework*, 2025.

when implementing them internally, reflecting an effort for greater legislative harmonization by the European Union. Thus, the distinctions between Directives and Regulations are, in practice, increasingly less rigid, reflecting a movement of approximation between the two instruments regarding normative standardization[13]. This movement is perfectly in line with the ideology of regulatory standardization of the New Approach and the NLF.

Entering specifically into its constitution, the New Legislative Framework (NLF) is essentially composed of Regulation (EC) No. 765/2008, Decision No. 768/2008/EC, and Regulation (EU) No. 2019/1020, acts referred to as horizontal because they establish general rules of application[14].

In addition to these normative instruments, the relevance of the document "Evaluation of the New Legislative Framework" stands out, which provides a comprehensive assessment of the NLF's effectiveness. A joint reading of these materials allows for a deeper and more systematic understanding of the regulatory structure that supports the European model of technical harmonization in favor of the formation of a single market.

Some of the directives and regulations, besides the AI Act, that compose the NLF are, for example: Toys' safety (Directive 2009/48/EC); Electrical equipment designed for use within certain voltage limits (Directive 2014/35/EU); Machinery(Directive 2006/42/EC); Cableway installations (Regulation (EU) 2016/424); Radio equipment (Directive 2014/53/EU); Medical devices (Regulation (EU) 2017/745), among others. These are called sectoral harmonization acts of the Union[15].

Certain acts provide for the exclusion from their scope of application of products covered by other legislations, avoiding the overlapping of requirements. On the other hand, if there is no such provision, the principle of simultaneous application is resorted to, where the product must comply with all acts that concern it[16]. At this point lies one of the main criticisms of this regulatory model, the overlapping of norms,

---

13 DU BOISPÉAN; MUECK; GAIE, *Introduction to the European New Legislative Framework*, 2025.

14 EUROPEAN COMMISSION, *Blue Guide*, 2022.

15 EUROPEAN COMMISSION, *Blue Guide*, 2022.

16 EUROPEAN COMMISSION, *Blue Guide*, 2022.

which can generate excessively complex scenarios for regulated subjects, sometimes with disproportionate burdens[17].

For these reasons it is important to emphasize that the AI Act should not be understood as a self-applicable and exhaustive legal instrument. Several of its provisions require additional technical development, and harmonized standards, although formally non-binding, are essential for the concretization and operationalization of the legal requirements provided in the regulation. Furthermore, as it is a technology with broad and varied application, its regulation, in practice, can attract the incidence of a series of legislative productions.

4.1. Regulation 765/2008, Regulation 2019/1020 and Decision No. 768/2008/EC

Before moving on to the analysis of each of the regulations, it should be noted that Regulation 765/2008 was amended, but not repealed, by Regulation 2019/1020. The latter is defined as a special law, considering that it applies to all products that do not have similar inspection requirements in their markets of origin, for example, it establishes measures especially aimed at products imported from non-EU countries through online sales.[18].

Regulation (EC) No. 765/2008 aims to establish a balance between two fundamental principles of European integration, the free movement of goods within the internal market and the protection of relevant public interests, such as individual health and safety[19]. To this end, the regulation adopts a set of horizontal provisions,

---

[17] EBERS, Martin. Truly Risk-based Regulation of Artificial Intelligence: How to Implement the EU's AI Act. *European Journal of Risk Regulation*. Cambridge: Cambridge University Press. p. 01-20, 2024. Available at: <https://www.cambridge.org/core/journals/european-journal-of-risk-regulation/article/truly-riskbased-regulation-of-artificial-intelligence-how-to-implement-the-eus-ai-act/E526C1D0D7368F9691082220609D60F4>. Access: 21 Jan. 2025.

[18] DU BOISPÉAN; MUECK; GAIE, *Introduction to the European New Legislative Framework*, 2025.

[19] EUROPEAN UNION. Regulation (EC) No 765/2008 of the European Parliament and of the Council of 9 July 2008 setting out the requirements for accreditation and market surveillance relating to the marketing of products and repealing Regulation (EEC) No 339/93. *Official Journal*

applicable to various sectors and types of products, with the aim of standardizing accreditation and market surveillance criteria. By harmonizing these procedures, the aim is to promote the safe cross-border movement of goods.

As it is a horizontal legislation, it is important that the norm unifies some fundamental concepts, which is why in its Article 2 it brings a list of definitions of concepts such as "Manufacturer"; "Importer"; "Distributor". Paul Voigt and Nils Hullen clarify that although sectoral acts, such as the AI Act, must follow horizontal harmonization rules, they may have their own peculiarities, always respecting similarity with the general rule. For example, in the act regulating artificial intelligence systems, the manufacturer is considered a provider[20].

The main contributions brought by this regulation for the pursuit of the consolidation of the European single market were the accreditation system; the market surveillance framework; the control of products entering the European Union; and the CE marking[21][22].

Accreditation is carried out by an official body, usually from each Member State. In this procedure, it is certified that a conformity assessment body has the technical competencies for the function, issuing the respective certificate[23]. Before the entry into force of Regulation (EC) No. 765/2008, the absence of common accreditation rules among Member States resulted in a fragmented scenario, in which certificates issued by national bodies were not necessarily uniformly recognized by other economic operators or public authorities of the European Union. This led to the need for multiple accreditations for the same conformity assessment body, which, in turn, significantly increased operational costs for both companies and the assessment bodies themselves[24]. Thus, a duly accredited body can carry out conformity assess-

of the European Union, 13 August 2008. Available at: https://eur-lex.europa.eu/legal-content/PT/TXT/?uri=CELEX%3A32008R0765. Accessed: 17 June 2025.

20 VOIGT, Paul; HULLEN, Nils. *The EU AI Act: Answers to Frequently Asked Questions*. New York: Springer Nature, 2024.

21 EUROPEAN UNION, *Regulation (EC) No 765/2008*, 2008.

22 EUROPEAN COMMISSION, *Blue Guide*, 2022.

23 EUROPEAN UNION, *Regulation (EC) No 765/2008*, 2008.

24 EUROPEAN COMMISSION, *Blue Guide*, 2022.

ment.

The said conformity assessment is provided for in Decision No. 768/2008/EC. It is divided into several modules (A to H, with their variants), and in each of them, the norm establishes duties to be fulfilled by the manufacturer[25]. In module A, for example, the manufacturer must only declare that the duties of technical documentation and manufacturing rules are being complied with (self-declaration) and must carry out the conformity marking. In module H, on the other hand, a notified body intervenes in the phases prior to production and during it[26]. The legislator then selects, from these modules, the most suitable for the specific needs of a given sector. "Less burdensome" modules should be selected taking into account the type of products and the risks involved. This seeks to limit the number of possible conformity procedures while the legislator's power of choice ensures covering the widest possible range of products[27].

This point is crucial for understanding the *raison d'être* of the AI Act. Understanding its underlying logic requires recognizing its insertion into the context of the New Legislative Framework (NLF)[28]. Its structure based on risk classification: unac-

---

25 Depending on the legislator, conformity assessment can be carried out solely by the manufacturer, without any third-party intervention; with the intervention of an accredited internal body, which is part of the manufacturer's organization but performs this activity impartially; or externally, an independent third party. Furthermore, it is important to highlight that the distinction between conformity assessment bodies and "notified bodies" lies in the fact that the latter correspond to those formally designated by a competent national authority and notified to the European Commission and other Member States, with a view to carrying out conformity assessment procedures in cases where the legislation requires the intervention of an independent third party.. (EUROPEAN COMMISSION, *Blue Guide*, 2022).

26 EUROPEAN UNION. Decision No 768/2008/EC of the European Parliament and of the Council of 9 July 2008 on a common framework for the marketing of products and repealing Decision 93/465/EEC. *Official Journal of the European Union*, 13 August 2008. Available at: https://eur-lex.europa.eu/legal-content/PT/TXT/?uri=CELEX%3A32008D0768. Accessed: 17 June 2025.

27 EUROPEAN COMMISSION, *Blue Guide*, 2022.

28 AI Act Recital 9 Harmonised rules applicable to the placing on the market, the putting into service and the use of high-risk AI systems should be laid down consistently with Regulation (EC) No 765/2008 of the European Parliament and of the Council, Decision No 768/2008/EC of the European Parliament and of the Council and Regulation (EU) 2019/1020 of the European Parliament and of the Council (New Legislative Framework).

ceptable risk, high risk, transparency risk, and minimal risk is not merely accidental [29]. It is a strategy consistent with the European regulatory ecosystem focused on product safety, where the intensity of regulation is directly proportional to the risk involved. Among the main normative instruments that underpin this model is Decision No. 768/2008/EC, which establishes differentiated conformity assessment procedures, precisely based on the nature and severity of the risks presented by a given product[30].

Regarding the market surveillance framework and the control of products entering the European Union, returning to Regulation (EC) No. 765/2008, these provisions were replaced by the provisions of Regulation (EU) 2019/1020[31]. Therefore, it is appropriate to analyze the new provisions in force.

According to Regulation (EU) 2019/1020, each Member State must designate one or more market surveillance authorities in its territory. Among the powers of these authorities are the possibility of requesting documents and product samples from economic operators (including under false identity), carrying out inspections, applying sanctions, adopting corrective measures, among others[32]. It is important to mention that the actions of these authorities are also guided by risk. Article 19 establishes, for example, that products presenting a serious risk must be withdrawn or recalled, if there are no other effective means to eliminate it, or that their placement

---

29 VOIGT; HULLEN, *The EU AI Act*, 2024.

30 AI Act Article 29: Application of a Conformity Assessment Body for Notification 2. The application for notification shall be accompanied by a description of the conformity assessment activities, the conformity assessment **module or modules** and the types of AI systems for which the conformity assessment body claims to be competent, as well as by an accreditation certificate, where one exists, issued by a national accreditation body attesting that the conformity assessment body fulfils the requirements laid down in Article 31. Any valid document related to existing designations of the applicant notified body under any other Union harmonisation legislation shall be added.

31 EUROPEAN COMMISSION, *Blue Guide*, 2022.

32 EUROPEAN UNION. Regulation (EU) 2019/1020 of the European Parliament and of the Council of 20 June 2019 on market surveillance and compliance of products and amending Directive 2004/42/EC and Regulations (EC) No 765/2008 and (EU) No 305/2011. *Official Journal of the European Union*, 25 June 2019. Available at: https://eur-lex.europa.eu/legal-content/PT/TXT/?uri=CELEX%3A32019R1020. Accessed: 17 June 2025.

on the market be prohibited[33]. Therefore, once again, a direct dialogue is established with the risk-based regulation of the AI Act.

The aim of consolidating a single European market becomes more evident when the norm establishes a rapid information exchange system (RAPEX), in case a market surveillance authority intends to recall a certain irregular product and considers that the reasons for the measures or their respective effects extend beyond the territory of its Member State[34].

Regarding the control of products entering the European Union, a matter also updated by Regulation (EU) 2019/1020, it is established that each Member State must designate customs authorities to monitor products entering the common European market[35]. Once again, the model of this surveillance is based on a risk analysis, similar to that carried out by market surveillance authorities[36].

Moreover, the efforts for the consolidation of a single market are evident in Article 29 of Regulation (EU) 2019/1020, which creates the European Union network for product compliance, in which each Member State designates a representative to act, cooperatively, in customs surveillance[3738].

Having exposed some of the provisions of Regulation (EC) No. 765/2008: the accreditation system, related to the conformity assessment procedure detailed by Decision No. 768/2008/EC; the market surveillance framework and the control of products entering the European Union, amended by Regulation (EU) 2019/1020; it remains to analyze the CE marking, which is regulated by both Regulation (EC) No. 765/2008 and Decision No. 768/2008/EC.

While Regulation (EC) No. 765/2008 establishes the definition, format, and general principles governing the CE marking. Decision No. 768/2008/EC provides for conformity assessment procedures leading to its affixing[39].

---

33 EUROPEAN UNION, Regulation (EU) 2019/1020, 2019.

34 EUROPEAN UNION, Regulation (EU) 2019/1020, 2019.

35 EUROPEAN UNION, Regulation (EU) 2019/1020, 2019.

36 Articles 25.3 and 11 3. Regulation (EU) 2019/1020.

37 EUROPEAN UNION, Regulation (EU) 2019/1020, 2019.

38 EUROPEAN COMMISSION, *Blue Guide*, 2022.

39 EUROPEAN COMMISSION, *Blue Guide,* 2022.

The CE marking is a symbol that the manufacturer affixes to the product to attest its conformity with the applicable essential requirements of the legislation[40]. This symbol must be visible, legible, and must be affixed before the product is placed on the market[41]. Evidently, as part of the NLF, the AI Act subjects AI systems to these marking rules.[42]

The analysis of the main horizontal norms that make up the New Legislative Framework (NLF) allows for three fundamental conclusions. Firstly, these norms have as their central objective the unification of the European Union's internal market, by establishing common rules applicable to all Member States, as well as by instituting cooperation mechanisms aimed at supervising the movement of products. Secondly, the regulatory model is primarily based on risk analysis, which is concretized, for example, in the application of the conformity assessment modules provided for in Decision No. 768/2008/EC, as well as in the provisions for controlling the entry of products into the Union's territory, as regulated by Regulation (EU) 2019/1020. Finally, considering that the AI Act was conceived within the scope of the NLF, it is essential that its normative structure is aligned with the principles, procedures, and guidelines established by the normative acts that integrate this framework, thus ensuring the coherence and effectiveness of the European product safety system.

4.2. Evaluation of the New Legislative Framework

Years after the implementation of the NLF, the European Commission conducted a study to evaluate its performance between 2014 and 2020[43]. For the evaluation of the NLF's performance, its specific objectives were taken into account: 1) Reinforce

---

40 EUROPEAN UNION, Regulation (EC) No 765/2008, 2008.

41 EUROPEAN UNION, Decision No 768/2008/EC, 2008.

42 AI Act Article 48: CE Marking 1. The CE marking shall be subject to the general principles set out in Article 30 of Regulation (EC) No 765/2008.

43 EUROPEAN UNION. *European Commission. Commission Staff Working Document – Evaluation of the New Legislative Framework.* SWD(2022) 364 final/2, Corrigendum. Brussels, 16 November 2022. Available at: https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/12654-Industrial-products-evaluation-of-the-new-legislative-framework_en. Accessed: 19 June 2025.

the New Approach model; 2) Support the consistency and coherence of European Union harmonization legislation; 3) Strengthen the conformity assessment system; and 4) Reinforce the clarity and credibility of the CE marking[44].

Initially, authorities from Member States, companies and entrepreneurs, economic operators, such as distributors and importers, consumer organizations, and EU citizens were consulted[45]. 125 contributions were received, most coming from business associations (29.6%), followed by companies or trade organizations (20%) and EU citizens (22.4%). Public authorities accounted for 16% of responses, while other categories totaled 8.8%. Participation from non-EU citizens accounted for 1.6%, and both consumer organizations and non-governmental organizations (NGOs) contributed only 0.8% each[46].

A summary report containing the responses of economic agents was prepared, with the main objective of evaluating the effectiveness of the New Legislative Framework (NLF) in standardizing regulatory tools aimed at the circulation of products, in establishing a system for accreditation, and in consolidating rules related to the CE marking[47]. It is noteworthy that this summary report reflects the opinions of the involved economic agents, and not of the European Commission itself; however, its results subsidized the elaboration of the final report.

Among the twenty-two questions proposed in the summary report, one can be highlighted: "How effective has the NLF been in strengthening a high level of protection of public interests, such as health and safety, consumer protection, and environmental preservation?" Of the 125 responses, 47 (37.6%) answered "very effective"; 36

---

44 EUROPEAN COMMISSION, *Blue Guide*, 2022.

45 EUROPEAN COMMISSION. *Evaluation and Fitness Check Roadmap* – Ares(2020) 6353387. Brussels: European Commission, 2020. Available at: https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/12461-Evaluation-of-the-New-Legislative-Framework-for-products_en. Accessed: 17 June 2025.

46 EUROPEAN COMMISSION. *Factual summary report on the public consultation on the New Legislative Framework (NLF)*. Ares(2022) 2416012, March 2022. Available at: https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/12654-Evaluation-of-the-New-Legislative-Framework-for-EU-legislation-on-industrial-products/public-consultation_en. Accessed: 17 June 2025.

47 EUROPEAN COMMISSION, NLF *Public Consultation Report Summary,* 2022.

(28.8%) answered "somewhat effective"; and 30 did not respond[48]. Thus, of those who responded (95), 83 (87.4%) recognized the effectiveness of the NLF in this regard.

In another question: "How effective has the NLF been in strengthening the free movement of products in the single market?" Of the 125 responses, 52 (41.6%) answered "very effective"; 34 (27.2%) answered "somewhat effective"; and 30 also did not respond[49]. Thus, of those who responded (95), 86 (90.5%) recognized the effectiveness of the NLF in this regard.

These examples show that, in a preliminary analysis of the report, the New Legislative Framework (NLF) received a largely positive reception from economic agents. This initial perception was later corroborated by the final report published by the European Commission.

Regarding this latter report, it is interesting to note that it was prepared in 2022, while the AI Act was still a legislative proposal, but even so, the evaluation conducted by the European Commission emphasized the importance for regulatory advances within the NLF aimed at what is called "digital transition". In relation to the relevance of the NLF, it is concluded, in line with the summary report, that its objectives and principles continue to be aimed for the future, especially in view of the phenomenon of digitalization[50].

Given the above, it is observed that the New Legislative Framework (NLF) has been positively evaluated by civil society, economic operators, and the European Union Administration itself. Its relevance stems from the central role it plays in promoting the free movement of products within the European internal market, while establishing mechanisms aimed at ensuring the safety of these products. This widespread acceptance and its regulatory effectiveness can help explain the European legislator's option to frame new normative acts, such as the Artificial Intelligence Act, within the NLF structure, ensuring its compatibility with the principles of the New Approach and its coherence with the current legal framework.

---

48 EUROPEAN COMMISSION, NLF *Public Consultation Report Summary,* 2022.

49 EUROPEAN COMMISSION, NLF *Public Consultation Report Summary,* 2022.

50 EUROPEAN COMMISSION, SWD, 2022.

5. ARTIFICIAL INTELLIGENCE ACT X GENERAL DATA PROTECTION REGULATION

As Marco Almada and Anca Radu point out, the AI Act has been subject to significant criticism regarding its deficient safeguarding of fundamental rights. The main concern lies in the fact that, by being inserted into the normative framework primarily focused on product safety, in the context of the New Legislative Framework, the protection of individual rights is carried out only indirectly. Added to this is the logic adopted by the risk-based approach, which starts from the analysis of the probability of a harmful event occurring and the severity of its effects. This logic, the authors affirm, proves incompatible with the regime of fundamental rights, which, by their nature, cannot be subject to quantification and balancing[51].

Martin Ebers argues that there is a structural incongruity between the logic of risk-based regulation and the effective protection of fundamental rights[9]. According to the author, while the former admits "degrees of legality" according to risk assessment, the latter operates under a binary logic, where a certain conduct is legal or illegal, and fundamental rights are non-negotiable. Furthermore, he emphasizes that risk-oriented regulation aims to protect collective interests, whereas the protection of fundamental rights focuses on the individual sphere[52].

Although acknowledging the challenges of this conciliation, Ebers observes with concern the AI Act's attempt to expand the scope of risk-based regulation to also cover the fundamental rights of citizens[53][54]. Such ambition, according to the author,

---

[51] ALMADA, Marco; RADU, Anca. *The Brussels Side-Effect: How the AI Act Can Reduce the Global Reach of EU Policy*. German Law Journal. German Law Journal. Frankfurt: German Law Journal e.V. v. 25, n. 4, p. 646-663, June, 2024.

[52] EBERS, *Truly Risk-based Regulation of Artificial Intelligence*, 2024.

[53] EBERS, *Truly Risk-based Regulation of Artificial Intelligence*, 2024.

[54] Article 27: "Fundamental Rights Impact Assessment for High-Risk AI System". "Article 1. The purpose of this Regulation is to *improve the functioning of the internal market* and promote the uptake of human-centric and trustworthy artificial intelligence (AI), while ensuring a high level of protection of health, safety, *fundamental rights* enshrined in the Charter, including democracy, the rule of law and environmental protection, against the harmful effects of AI systems in the Union and supporting innovation" (Europe Union, 2024).

extrapolates the scope of the European Union's competence outlined by Article 114 of the Treaty on the Functioning of the European Union (TFEU), which grants the Union competence only to remove barriers to the free movement of products in the internal market[55].

Notwithstanding the AI Act's attempt to include the protection of fundamental rights in its scope, a closer analysis reveals that its primary focus is on product safety. This normative orientation becomes even more evident when comparing the AI Act with legislations such as the General Data Protection Regulation (GDPR).

5.1. Artificial Intelligence Act

The purpose of this topic is not to delve into the normative intricacies of the AI Act, but rather to highlight that its regulatory structure is primarily oriented towards product safety and the consolidation of the European internal market, in the mold of the New Legislative Framework (NLF). Thus, although the AI Act mentions the protection of fundamental rights, this occurs indirectly and accessorily, being more a desirable consequence of technical regulation than a central objective of the diploma.

First, regarding the object under regulation, the act applies to artificial intelligence systems and general-purpose AI (GPAI) models. Voigt and Hullen explain the difference between AI systems and models: AI models are necessary components of AI systems. AI models, when added to other elements, such as a user interface, form an AI system[56].

This is relevant to understanding the AI Act's insertion into the NLF context. By largely directing its regulation to AI systems and exceptionally to GPAI models (a type of AI model), the regulation concentrates its scope on artificial intelligence products at an advanced stage of development, ready for circulation in the European Union's internal market. These products, by already having a user interface, are

---

[55] EUROPEAN UNION. Treaty on the Functioning of the European Union. Lisbon, 13 December 2007. Available at: https://eur-lex.europa.eu/legal-content/PT/TXT/?uri=CELEX%3A12012E%2FTXT. Accessed: 19 June 2025.

[56] VOIGT; HULLEN, *The EU AI Act*, 2024.

capable of being made available to end-users, which justifies their inclusion in the NLF logic.

Another factor that reinforces this line of argument is the explicit exclusion, by the AI Act itself, of artificial intelligence systems used outside the consumer context, such as those intended for military, defense, and national security purposes, as well as for scientific research and development activities[57].

Next, in Article 3, economic operators are defined: provider, product manufacturer, deployer, authorized representative, importer, and distributor [58], showing a clear resemblance to the descriptions contained in the acts that compose the NLF [59]. The same logic applies to concepts such as "placing on the market" and "making available on the market"[60]. These adopted terminologies do not directly refer to the language of fundamental rights or individual protection, but rather to the logic of the supply chain, which highlights their close relationship with regulation focused on consumer law, the consolidation of a product safety system, and the European internal market.

The act is structured based on risk classification: unacceptable risk, high risk, transparency risk, and minimal risk[61]. Depending on the use attributed to the artificial intelligence system or the context of its application, it falls into one of these categories. The consequence of this is that, depending on the classification given to the system, economic operators will have different types of legal obligations. Among these duties, most are directed to operators of high-risk systems, mainly providers and deployers.[62]

---

57 EUROPEAN UNION, Regulation (EU) 2024/1689 (Artificial Intelligence Act), 2024.

58 EUROPEAN UNION, Regulation (EU) 2024/1689 (Artificial Intelligence Act), 2024.

59 AI Act, Article 3 (6) 'importer' means a natural or legal person located or established in the Union that places on the market an AI system that bears the name or trademark of a natural or legal person established in a third country / Regulation (EC) No 765/2008, Article 2 (5) 'importer' shall mean any natural or legal person established within the Community who places a product from a third country on the Community market.

60 EUROPEAN UNION, Regulation (EU) 2024/1689 (Artificial Intelligence Act), 2024.

61 VOIGT; HULLEN, *The EU AI Act*, 2024.

62 FAZLIOGLU, Müge. E*U AI Act Compliance Matrix. IAPP – International Association of Privacy Professionals*, 2024. Available in: https://iapp.org/media/pdf/resource_center/eu_ai_act_compli-

The following obligations can be cited, implementation of a risk management system (Article 9); data governance (Article 10); elaboration of adequate technical documentation (Article 11); maintenance of records on system operation (Article 12); ensuring transparency and providing information to users (Article 13); implementation of a quality management system (Article 17); cooperation with legal authorities (Article 21); conformity assessment (Article 43); issuance of the declaration of conformity (Article 47); CE marking (Article 48); formal registration of operators and systems (Article 49); post-market monitoring (Article 72); obligation to report serious incidents to competent authorities (Article 73)[63]. Non-compliance with certain provisions of the AI Act can result in fines ranging from 7.5 million euros, 15 million euros, and, in the most serious cases, up to 35 million euros[64].

Several provisions of the AI Act directly interact with the mechanisms structured within the scope of the New Legislative Framework (NLF), as analyzed in the previous chapter. The definition of economic operators, the adoption of a normative classification based on risk levels, and the distribution of legal responsibilities according to these two elements reveal the conformation of a legal-regulatory ecosystem oriented towards product safety, in this case, artificial intelligence systems, and ensuring their free movement among the European Union Member States.

This observation is not limited to the AI Act but extends to the directives and regulations that compose the NLF, which adopt a structure similar to that mentioned above. For example, the regulation of medical devices, in its Article 51, establishes the classification of devices into classes I, IIa, IIb, and III, considering their intended use and the risks they present, according to the rules set out in Annex VIII[13]. Similarly, the classification attributed to the device implies different obligations for the economic operators involved[65]. Therefore, the risk-based regulation of the AI Act is not innovative. Despite appearing emblematic to the rest of the world, it is simply an

ance_matrix.pdf. Access in: 19 June 2025.

63 EUROPEAN UNION, Regulation (EU) 2024/1689 (Artificial Intelligence Act), 2024.

64 VOIGT; HULLEN, *The EU AI Act*, 2024.

65 EUROPEAN UNION. Regulation (EU) 2017/745 of the European Parliament and of the Council of 5 April 2017 on medical devices, and repealing Directives 90/385/EEC and 93/42/EEC. Official Journal of the European Union, 5 May 2017. Available at: https://eur-lex.europa.eu/legal-content/PT/TXT/?uri=CELEX%3A32017R0745. Accessed: 19 June 2025.

adaptation to the normative context, ensuring its coherence.

The AI Act indeed attributes scarce attention to the consecration of general principles and the provision of subjective rights. Voigt and Hullen highlight only a few provisions that address such elements. Recital 69, which mentions the principles of Privacy by Design and Privacy by Default, which, however, do not constitute legislative novelty, as they are already provided for in Recital 78 of the GDPR; the right to lodge a complaint with the market surveillance authority, as provided in Article 85; the right to obtain an explanation of a decision resulting from the use of a high-risk AI system, provided in Article 86; and the right to lodge a reasoned complaint with the AI Office, in cases of alleged violation of the regulation, as provided in Article 89(2)[66].

5.2. General Data Protection Regulation

In contrast to the AI Act, the General Data Protection Regulation (GDPR) deviates from the scope of the New Legislative Framework (NLF), as it does not concern product safety, but rather the protection of personal data. It is, therefore, a regulation aimed at protecting an intangible asset, not introducible into goods circulating in the European Union's internal market. For this reason, its legal legitimacy basis is not found in Article 114 of the Treaty on the Functioning of the European Union (TFEU), as is the AI Act's, but in Article 16(2) of the same treaty[67].

---

66 VOIGT; HULLEN, *The EU AI Act*, 2024.

67 Article 114. 1. 1. Save where otherwise provided in the Treaties, the following provisions shall apply for the achievement of the objectives set out in Article 26. The European Parliament and the Council, acting in accordance with the ordinary legislative procedure and after consulting the Economic and Social Committee, adopt the measures for the approximation of the provisions laid down by law, regulation or administrative action in Member States which have as their object the establishment and functioning of the internal market. Article 16. 2. The European Parliament and the Council, acting in accordance with the ordinary legislative procedure, shall lay down the rules relating to the protection of individuals with regard to the processing of personal data by Union institutions, bodies, offices and agencies, and by the Member States when carrying out activities which fall within the scope of Union law, and the rules relating to the free movement of such data. Compliance with these rules shall be subject to the control of independent authorities. (EUROPEAN UNION, Treaty on the Functioning of the European Union (TFEU), 2007).

In an analysis of the GDPR, Chapter 2 (Articles 5 to 11) establishes a series of principles related to the processing of personal data, such as the principle of adequacy, consent, transparency, all aimed at safeguarding fundamental rights, especially regarding the right to privacy. Furthermore, Chapter 3 (Articles 12 to 23) provides for the rights of data subjects whose data are processed, such as the right to access information about processing and their data, the right to rectification, erasure, objection, and data portability[68].

This article does not intend to individually delve into each of these provisions, but it is important to highlight that all of them converge towards a primary objective, the protection of fundamental individual rights. Perhaps this normative orientation can explain, at least in part, the wide international acceptance and influence of the regulation. Since its original conception (Directive No. 1995/46/EC), the GDPR has served as a reference for various foreign legislations, including the current Brazilian General Data Protection Law, which largely reproduces the principles and rights enshrined in European law[69-70].

The "Brussels Effect" (The Brussels Side-Effect) is the phenomenon by which normative acts transcend the European Union's borders, influencing regulatory standards and public policies on a global scale[71]. Although the GDPR is frequently cited as a paradigmatic example of this effect, raising debates about the possibility of the AI Act following a similar path, it is essential to recognize the differences in nature between these two legislations, as well as the distinct normative contexts in

---

68 EUROPEAN UNION. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data (General Data Protection Regulation – GDPR). Official Journal of the European Union, 4 May 2016. Available at: https://eur-lex.europa.eu/legal-content/PT/TXT/?uri=CELEX%3A32016R0679. Accessed: 10 June 2025.

69 PARENTONI, Leonardo. O direito ao esquecimento (Right to Oblivion). In: LUCCA, Newton De; SIMÃO FILHO, Adalberto; LIMA, Cíntia Rosa Pereira de (org.). *Direito & Internet III: Marco Civil da Internet* (Lei nº 12.965/2014). 1. ed. São Paulo: Quartier Latin, 2015. p. 539–618. Available at: https://www.researchgate.net/publication/299821661_O_Direito_ao_Esquecimento_Right_to_Oblivion. Access in: 10 Jue. 2025.

70 SCASSA, Teresa. Sovereignity and the Governance of Artificial Intelligence. *UCLA Law Review*. Los Angeles: University of California School of Law. v. 71, n. 1, p. 214-229, May, 2024.

71 ALMADA; RADU, *The Brussels Side-Effect*, 2024.

which they are inserted.

5.3. Conclusion

The Artificial Intelligence Act is embedded in a broader normative complex, aimed at consolidating the European single market through the establishment of common rules related to product safety. Criticisms that point to an alleged absence of principles and fundamental guarantees aimed at protecting the human person in the text of the AI Act generally disregard the normative and institutional context in which this regulation is inserted. Its primary objective is not the direct protection of fundamental rights, but rather the creation of a harmonized normative environment that favors the safe circulation of artificial intelligence systems in the European economic space.

From this perspective, the protection of fundamental rights figures in the background, but it is not ignored. The normative strategy adopted by the European legislator is inductive, by establishing minimum technical safety requirements for the commercialization of AI systems, it is presumed that, indirectly, certain aspects of consumer protection and human dignity will also be promoted.

In this sense, it is methodologically inadequate to compare the AI Act with normative instruments such as the General Data Protection Regulation (GDPR), without considering their respective original normative contexts. The GDPR is not part of the NLF logic, nor does it concern the safety of physical products, its field of incidence is different, the protection of personal data. By dealing with a legal asset of an intangible nature, which is not incorporated into a marketable product in traditional ways, data protection requires its own legal regime, distinct from that applied to the regulation of technologies incorporated into consumer goods, as is the case with AI under the AI Act.

On the other hand, the topic of artificial intelligence allows for multiple regulatory approaches. The European Union's option, at least at this moment, was to privilege a perspective centered on product safety and the protection of the internal market, perhaps herein lies a possible criticism.

This does not prevent, however, a specific norm from being elaborated in the future for the regulation of AI from the perspective of fundamental rights, in the

manner of the GDPR[72]. A legitimate and more forceful criticism, therefore, lies in the evaluation of the normative priority adopted. By privileging the economic and regulatory dimension of AI, the European Union would have, possibly, left a more directly individual rights-oriented approach in the background.

This path can be justified, in part, by the fear that excessively premature or restrictive regulation could inhibit the development of technology in the European space. However, it can be counter-argued that even currently existing regulation aimed at market safety and conformity imposes relevant costs and barriers on economic operators, which in itself could negatively impact the pace of technological innovation[73].

Furthermore, it can be argued that fundamental rights are not completely unprotected against the risks arising from AI. Various general normative diplomas, such as the GDPR itself, the Charter of Fundamental Rights of the European Union, or the New Product Liability Directive (Directive (EU) 2024/2853) which deals with civil liability for damages caused by defective products[74] remain fully applicable to AI systems.

In the Brazilian context, this discussion acquires additional contours, given the analytical and eclectic nature of the 1988 Constitution, which already enshrines an extensive list of fundamental rights. This characteristic raises the question of the real need for creating a specific normative framework for AI, or whether it would be more appropriate to interpret it in light of the existing legal order. It is, therefore, an open debate, with time, in this case, being a determining element to evaluate the sufficiency or insufficiency of the currently available normative instruments.

---

[72] Clearly, the legal basis for this depends on a provision in the Treaty on the Functioning of the European Union (TFEU).

[73] It is estimated that the compliance rules imposed by the AI Act, including conducting compliance assessments and implementing quality management systems, generate costs for companies that provide and/or use artificial intelligence technologies ranging from €193,000 to €330,000 (GIKAY, 2024).

[74] EUROPEAN UNION. Directive (EU) 2024/2853 of the European Parliament and of the Council of 23 October 2024 on liability for defective products and repealing Council Directive 85/374/EEC. Official Journal of the European Union, L, 18 November 2024. Available at: https://eur-lex.europa.eu/eli/dir/2024/2853/oj. Accessed: 30 June 2025.

6. BRAZILIAN BILL 2,338/23

Bill No. 2,338, of 2023, currently under deliberation in the National Congress, constitutes the main legislative proposal aimed at regulating artificial intelligence systems in Brazil. The proposition adopts a risk-oriented regulatory approach, simultaneously based on the protection of fundamental rights[75].

Chapter III of the normative proposition addresses the "categorization of risks." According to Article 13, it is incumbent upon the provider, prior to making the system available on the market or using it in services, to carry out, document, and register a preliminary assessment aimed at classifying the degree of risk of the artificial intelligence system. This classification, however, may be reviewed by determination of the competent authority. In Section II of the said chapter, AI systems classified as "excessive risk" are listed, with Article 14 providing for cases of prohibited uses of AI systems. Article 17, in turn, deals with the categorization of high-risk AI systems.

The proposed regulatory structure establishes that the duties of operating agents vary in conformity with the level of risk attributed to the system. Chapter IV, especially its Section I, provides for general governance duties, applicable to all AI systems. Section II contemplates additional specific governance measures for high-risk systems. In Section III, the mandatory algorithmic impact assessment is established, intended for systems classified as high-risk based on the preliminary analysis.

Parallel to the focus on risk management, the normative project also expressly dedicates itself to the consecration of principles and the safeguarding of fundamental rights. This is evidenced in Articles 2 and 3, which enshrine principles such as "centrality of the human person" and "non-discrimination." Complementarily, Chapter II, titled "Rights" (Arts. 5 to 12), establishes guarantees such as the "right to explana-

---

75 BRAZIL. Bill No. 2,338 of 2023 (PL 2338/2023). Provides for the development, promotion and ethical and responsible use of artificial intelligence based on the centrality of the human person. Author: Senator Rodrigo Pacheco (PSD/MG). Introduced in the Plenary of the Federal Senate on 3 May 2023; approved by the Plenary on 10 December 2024; transmitted to the Chamber of Deputies on 17 March 2025. Available at: https://www25.senado.leg.br/web/atividade/materias/-/materia/157233. Accessed: 22 June 2025.

tion," the "right to privacy," and the "right to contest automated decisions." Furthermore, Chapter V addresses civil liability, foreseeing the possibility of redress for damages resulting from the violation of fundamental rights.

Bill No. 2,338/2023 adopts a risk assessment model, inspired by the European regulation. However, it distinguishes itself by attempting to harmonize this approach with the direct protection of fundamental rights, despite the criticisms already raised regarding the possibility of this reconciled form of regulation.

When analyzing Brazilian General Data Protection Law,[76] a notable convergence with the General Data Protection Regulation (GDPR) can be identified. However, in the case of the AI Act, this convergence is not as pronounced, with the exception of the adoption of the transversal risk-based approach. Such a distinction, from this author's point of view, represents an advancement, given that the GDPR was developed in a normative environment with greater institutional freedom, while the AI Act is conditioned by the New Legislative Framework (NLF), which imposes some predefined regulatory structures.

Regarding the adoption of the risk categorization system, it is closely related to the provisions foreseen by the norms that make up the NLF, which raises questions about the advisability of its replication in the Brazilian legal system. This work does not intend to state whether risk-based regulation constitutes the best strategy to deal with the challenges of artificial intelligence, as this answer depends on a broader and more consolidated implementation of norms at a global level. The central question sought to be investigated is: when elaborating Bill No. 2,338/2023, did Brazilian legislators adequately consider the legal context of origin of the European AI Act?

A possible line of argument favoring an affirmative answer sustains that the Brazilian legislator did not merely replicate the European model but went further by expressly prioritizing the protection of fundamental rights.

On the other hand, the incorporation of risk-based classification, in the mold of the AI Act, raises doubts about the originality and adequacy of the Brazilian model.

---

[76] BRAZIL. Law No. 13,709 of 14 August 2018. Provides for the protection of personal data and amends Law No. 12,965 of 23 April 2014 (Internet Civil Framework Act). Official Gazette of the Union: Section 1, Brasília, DF, 15 August 2018, p. 1. Available at: https://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/L13709.htm. Accessed: 22 June 2025..

Thus, a second question arises: does risk-based regulation, in fact, constitute the best paradigm to face the regulatory challenges posed by artificial intelligence systems, or did the national legislator merely reproduce, uncritically, a foreign model alien to its specific normative context?

The answer to this question demands in-depth studies, strongly encouraged by the present author, aimed at discourse analysis applied to the legislative process. For now the question remains whether are the contexts of the European Union and Brazil so diferente. If so, would it not be beneficial to import the entire NLF?

7. THE PRODUCT SAFETY SYSTEM IN BRAZIL

In this chapter, it will be analyzed the possibility of importing the normative context in which the AI Act is embedded, that is, the New Legislative Framework of the European Union, as a solution for an eventual "decontextualized Brussels Effect," either through Bill No. 2,338/2023, currently under discussion, or other normative instruments that may be proposed. However, before that, it is necessary to examine whether, in fact, the Brazilian scenario significantly distances itself from that which shapes the European experience.

The investigation will focus on the Brazilian system for the safety of products circulating in the internal market. Law 5,966/73 created the National System of Metrology, Normalization, and Industrial Quality, with the purpose of formulating and executing the national policy of metrology, industrial normalization, and certification of industrial product quality[77]. In this system, Conmetro acts as the normative body and Inmetro as the executive body, both linked to the executive branch and regulated by Law 9,933/99.[78]

This law establishes in its Article 2 § 1 that Conmetro's technical regulations must address technical characteristics of inputs, final products, and services that do not fall

[77] BRAZIL. Law No. 5,966 of 11 December 1973. Establishes the National System of Metrology, Standardization and Industrial Quality, and provides for other matters. Official Gazette of the Union: Section 1, Brasília, DF, 13 December 1973.

[78] BRAZIL. Law No. 9,933 of 20 December 1999. Provides for the powers of Conmetro and Inmetro, establishes the fee for metrological services, and provides for other matters. Official Gazette of the Union: Section 1, Brasília, DF, 21 December 1999.

within the competence of other bodies and entities of the Federal Public Administration, regarding aspects related to safety, prevention of misleading commercial practices, protection of human, animal, and plant life and health, and the environment. These technical regulations must consider, where applicable, the content of technical standards adopted by the Brazilian Association of Technical Standards. Similarly, Article 3 establishes as Inmetro's functions to exercise administrative police power, issuing technical regulations in the areas of conformity assessment of products, inputs, and services, provided that they do not constitute the object of the competence of other bodies or entities of the federal public administration (item IV).

Both Inmetro and Conmetro, when participating in the elaboration of technical standards, must respect the competencies attributed to other regulatory institutions, especially sectoral regulatory agencies and, in certain circumstances, the Brazilian Association of Technical Standards (ABNT) itself. This is because, when an ABNT technical standard is expressly referenced by a regulatory normative act, it acquires a binding character. This is, therefore, a markedly fragmented regulatory scenario, especially when compared to the model adopted by the European Union.

In that context, greater institutional centrality is observed, with coordinated action by the European Parliament and the European Commission in formulating Directives and Regulations that establish the essential technical requirements for product safety. Complementary technical standards, which generate a presumption of conformity but are not mandatory, are left to specialized technical bodies, such as CEN, CENELEC, or ETSI. In this model, the normative provisions containing the essential requirements tend to present greater normative stability, while peripheral technical standards, of a non-mandatory nature, can be adapted with greater flexibility in the face of technological progress.

In contrast, in a context where various products have their technical requirements exhaustively detailed by one or more regulatory bodies, there is a significant risk that such standards become obsolete in the face of technological evolution, prove to be conflicting with each other, or present practical application difficulties.

In a comparative analysis with the New Legislative Framework, Inmetro itself recognizes that Brazil could benefit from an approach analogous to that practiced in the European Union, which would have the potential to "contribute to a decisive

modernization of Inmetro's regulation and the establishment of a modern, agile, effective, and efficient regulatory model"[79].

From this author's perspective, this structural difference between regulatory models stems from a fundamental element: sovereignty. Considering that the European Union is composed of sovereign Member States, often with divergent regulatory interests, normative acts produced within the community scope must be limited to what represents an essential core consensually agreed upon among participating countries, which, in itself, already constitutes a relevant challenge. This institutional limitation contributes to the adoption of more concise normative provisions and greater clarity of the legal-regulatory framework. In order to mitigate eventual regulatory gaps, the European model allows for the elaboration of specific technical standards subsidiarily, which, although not mandatory, generate a presumption of conformity when observed. In these hypotheses, it is sufficient for the regulated agent to demonstrate that the characteristics of the product or process are aligned with the safeguarding of the essential requirement provided in the reference norm.

In contrast, in the Brazilian legal system, the main actors involved in the regulatory structure, such as Conmetro, Inmetro, and sectoral regulatory agencies, are linked to or subordinated to the ministries that comprise the Federal Administration. Thus, as a unitary sovereign subject, it exercises the prerogative of enacting normative acts aimed at achieving the objectives of the Federative Republic of Brazil. This configuration grants greater freedom to the exercise of regulatory competence, which results, on one hand, in a higher number of objects under regulation and the proliferation of normative acts; on the other hand, it can lead to overlapping norms, conflicts of competence, and uncertainties regarding the delimitation of attributions among the various regulatory entities.

In the case of the European Union, the NLF aims to balance the pretensions of countries endowed with sovereignty. In Brazil, on the other hand, this does not apply.

---

[79] BRAZIL. National Institute of Metrology, Quality and Technology. Study on the regulatory models of Europe and the United States of America, and on the influence of Industry 4.0 on the modernization of Inmetro's regulation. Rio de Janeiro: Inmetro, 2020, p. 46. Available at: https://www.gov.br/inmetro/pt-br/assuntos/regulamentacao/modelo-regulatorio-do-inmetro/03-relatorio-modelo-europeu-v2.pdf. Access: 30 June 2025.

However, it is noteworthy that the EU regulatory landscape influenced the legislative integration process of Mercosur, also composed of sovereign countries, but, unlike the European case, Mercosur members did not achieve practical results from regulatory convergence[80].

However, it is worth remembering that the Brazilian product safety system is not entirely alien to the European experience. It is interesting to highlight that, in the first years of its creation, Inmetro received significant technical knowledge transfer from the German National Metrology Institute (PTB)[19]. During the 1970s and 1980s, fifteen Inmetro technicians were sent to Germany to learn the language and receive training in PTB laboratories.[81].

It is still possible to identify some similarities between the contexts analyzed. In both, accreditation systems, conformity assessment, market surveillance actions, as well as conformity markings, are present, each with its specificity. By way of example, two statutes may be considered, Ordinance No. 302, which deals with toy safety in Brazil, and the Toy Safety Directive 2009/48/EC of the European Union, previously presented.

The Brazilian act establishes that toys, manufactured, imported, distributed, and commercialized in national territory, gratuitously or onerously, must be compulsorily subjected to conformity assessment, provided for in Annex II, through the certification mechanism (Article 7). After certification, they must be registered with Inmetro as a condition for authorizing the use of the conformity identification seal (Article 8 and Annex III). Toys are subject, throughout the national territory, to market surveillance actions carried out by Inmetro and public law entities linked to it by delegation agreement (Article 12). Annex I provides for technical requirements,

---

80 THORSTENSEN, Vera; ARIMA JUNIOR, Mauro Kiithi; MEGALE, Tiago Matsuoka. Conformity assessment in Brazil and Argentina: differences, similarities and convergence models (Avaliação da conformidade no Brasil e na Argentina: diferenças, similaridades e modelos de convergência). *Boletim de Economia e Política Internacional (BEPI)*, no. 26, Sept. 2019 / Apr. 2020. Available at: https://repositorio.ipea.gov.br/handle/11058/9819. Access: 30 June 2025.

81 ALVES, L. P. *Triangular Technical Cooperation And The Role Of Inmetro*. Austral: Brazilian Journal of Strategy & International Relations, [S. l.], v. 2, n. 4, 2022. DOI: 10.22456/2238-6912.40498. Available in: https://seer.ufrgs.br/index.php/austral/article/view/40498. Accessed: 30 June 2025.

such as "5.1.6 Toys must be designed in such a way that the risk of fire, mechanical damage, lack of care or failure of a component does not impair their safety", "5.1.8 Toys must not be toxic"[82]. Thus, from the analysis of the legal text alone, a procedural approximation with the normative mechanisms adopted within the European Union is evident.

Regarding conformity assessment in Brazil, it is detailed by Inmetro Ordinance No. 200/2021[83]. The act regulates the functioning of the certification procedure through the definition of different models, classified as 1a, 1b, 2, 3, 4, 5, and 6. Depending on the product under regulation, one or some of these models are applied as a basis for the elaboration of a specific conformity assessment program, similarly to the modules provided in the European system.

Depending on the applicable model, only initial evaluation and/or periodic evaluations of the product and/or production process are carried out, through self-declarations and/or third-party verification. In the case of toys, for example, the supplier can opt for models 1b, 2, or 5, the latter being more complex.

Joyce Antoniolli et al. emphasize that the difference between one model and another is the company's planning, considering the urgency and product launch schedule. In System 1b, product safety is verified after a batch production, with minimal possibilities for corrective action, while in System 5, product safety is treated from a systemic view of the entire process that will originate the product.[84] This system seems to be based more on the strategy adopted by the manufacturer than on

---

82 BRAZIL. National Institute of Metrology, Quality and Technology – INMETRO. Ordinance No. 302 of 12 July 2021. Approves the Technical Quality Regulation and the Conformity Assessment Requirements for Toys – Consolidated. Official Gazette of the Union, Section 1, p. 30–44, 14 July 2021. Available at: http://www.inmetro.gov.br/legislacao/rtac/pdf/RTAC002801.pdf. Access: 1 July 2025.

83 BRAZIL. National Institute of Metrology, Quality and Technology – INMETRO. Ordinance No. 200 of 29 April 2021. Approves the General Requirements for Product Certification (RGCP) – Consolidated. Official Gazette of the Union, Section 1, p. 35–50, 30 April 2021. Available at: https://inmetro.gov.br/legislacao/rtac/pdf/RTAC002753.pdf. Access: 1 July 2025.

84 ANTONIOLLI, Joyce Multi da Silva; ROMÃO JÚNIOR, José Marcos; ANTUNES NETO, Joaquim M. F. *Avaliação da conformidade: importância e desafios para a indústria de brinquedos*. *Prospectus*, Itapira: FATEC de Itapira "Ogari de Castro Pacheco", v. 5, n. 2, p. 526–527, jul./dez. 2023. Available in: https://doi.org/10.5281/zenodo.10065005. Accessed: 30 June 2025.

the assessment of intrinsic risks to the product in question.

Despite this, according to Chamusca et al., both Brazil and the European Union consider risk as a factor influencing the choice of the conformity assessment procedure. But they recognize that they do not clarify how this factor influences decisions on conformity assessment rules.[85] In addition, they also highlight the influence of the production process on the selection of assessment procedures, citing the example of Canada, where third-party conformity assessment, considered more rigorous compared to self-declaration, is predominantly adopted in sectors that historically do not have a consolidated tradition of high quality[86].

Also, the efficiency of post-market control can also directly influence how conformity assessment is structured. As Thorstensen et al. point out, this type of control is more intensely used, in an integrated and systematized way, in the European Union than in Brazil. As a consequence, in the European context, there is a greater prevalence of self-conformity assessment mechanisms. Given the robust post-market surveillance, European manufacturers have a significant regulatory incentive to ensure product quality on their own. In Brazil, the dominant model is one that prioritizes prior verification carried out by a third party, through bodies accredited by Inmetro, since, often, post-market surveillance is deficient[87].

These considerations made, the need for a more in-depth analysis of the criteria adopted in defining the assessment models is imposed. The effective relevance of the "risk" factor in the choice of these procedures remains unclear, and it is not possible, at present, to establish definitive conclusions about the correlation between the European and Brazilian regulatory models.

Notwithstanding eventual normative similarities between different jurisdictions, the preliminary hypothesis is that the concrete effectiveness of legislation can vary significantly due to practical factors often neglected by a strictly literal reading of the norm. Elements pointed out such as business planning, the stage of productive devel-

---

85 CHAMUSCA, Roberta; ASSAFIN, Danielle; SAMPAIO, Cristiane; MACHADO, Raphael. "Critérios para a certificação de produtos na regulamentação do Inmetro". *P2P e Inovação,* vol. 6, n.º 1, p. 49–69, 2019. p.56. DOI: 10.21721/p2p.2019v6n1.

86 CHAMUSCA et al., *Criteria for Product Certification in Inmetro Regulation*, p.58, 2019.

87 THORSTENSEN et al., *Conformity Assessment in Brazil and Argentina*, 2019/2020.

opment of certain sectors, and the effectiveness of regulation mechanisms after the product is placed on the market directly influence how the law is applied and interpreted. Even if legal provisions have similar wording in different jurisdictions, the concrete effects of their implementation tend to diverge due to the factual, economic, and productive contextual particularities of each legal system.

The importation of the New Legislative Framework proves to be more complex than it appears at first glance. Beyond the implications related to sovereignty, this process involves considering the specificities of the Brazilian reality.

8. CONCLUSION

Uncertainties remain regarding the origin and foundation of the adoption of the risk-based approach in Bill No. 2,338/2023. It cannot be affirmed, with precision, whether this choice results from a conscious, technical, and contextualized deliberation on the part of the Brazilian legislator or if it results, on the contrary, from the purely normative influence of the European AI Act.

On the other hand, it was demonstrated throughout the work that the emblematic risk classification structure adopted by European regulation is strongly linked to the New Legislative Framework (NLF), a normative framework that, in turn, represents the continuity of the principles instituted with the New Approach. Such guidelines aim at legislative harmonization among the Member States of the European Union regarding product safety, with a view to consolidating the European internal market. To this end, they establish common normative mechanisms that promote regulatory cooperation between countries and enable the free movement of goods, based on minimum safety requirements.

Given this primary economic objective, the protection of fundamental rights appears secondarily, as an indirect reflection of the provisions aimed at product safety. In contrast, the Brazilian legal system, through Bill No. 2,338/2023, signals the adoption of a mixed approach. It incorporates the European model of risk classification, and grants centrality to the protection of fundamental human rights.

This duality may suggest a conceptual inconsistency, since risk-based regulation operates with a balance between acceptable and unacceptable risks, weighing their effects and benefits for the collectivity as a whole, while regulation oriented towards

safeguarding fundamental rights is structured for individual protection, of each human being in their set of non-negotiable rights.

Without intending to exhaust the analysis of the compatibility between these approaches, nor to take a side as to the superiority of one model over the other, this article aims at a critical reflection on the risks of decontextualized normative importation. Beyond the differences related to the issue of sovereignty, pointed out as a distinguishing factor of the legal-normative framework aimed at product safety, it is important to highlight that, even in the face of similarities observed between the Brazilian and European contexts regarding the procedures adopted for conformity assessment, there are still indications of differences in their application.

Therefore, there is concern about the possibility of reproducing the so-called Brussels Effect in relation to the AI Act, similar to what has already been observed with the General Data Protection Regulation. Considering that both norms are inserted in factually distinct contexts, it is essential that the Brazilian legislator adopts a critical stance when importing elements of the European model, especially regarding the adoption of risk classification as the structuring axis of artificial intelligence regulation in the country, particularly given the already observed tendency to incorporate this criterion as a central foundation in Bill No. 2,338/2023.

REFERENCES

ALMADA, Marco; RADU, Anca. *The Brussels Side-Effect: How the AI Act Can Reduce the Global Reach of EU Policy*. German Law Journal. German Law Journal. Frankfurt: German Law Journal e.V. v. 25, n. 4, p. 646-663, June, 2024.

ALVES, L. P. *Triangular Technical Cooperation And The Role Of Inmetro*. AUSTRAL: Brazilian Journal of Strategy & International Relations, [S. l.], v. 2, n. 4, 2022. DOI: 10.22456/2238-6912.40498.

ANTONIOLLI, Joyce Multi da Silva; ROMÃO JÚNIOR, José Marcos; ANTUNES NETO, Joaquim M. F. *Avaliação da conformidade: importância e desafios para a indústria de brinquedos*. *Prospectus*, Itapira: FATEC de Itapira "Ogari de Castro Pacheco", v. 5, n. 2, p. 526–527, jul./dez. 2023.

BRAZIL. Bill No. 2,338 of 2023 (PL 2338/2023). Provides for the development, promotion and ethical and responsible use of artificial intelligence based on the centrality of the human person. Author: Senator Rodrigo Pacheco (PSD/MG). Introduced in the Plenary of the Federal Senate on 3 May 2023; approved by the Plenary on 10 December 2024; transmitted to the Chamber of Deputies on 17 March 2025.

BRAZIL. Law No. 13,709 of 14 August 2018. Provides for the protection of personal data and amends Law No. 12,965 of 23 April 2014 (Internet Civil Framework Act). Official Gazette of the Union: Section 1, Brasília, DF, 15 August 2018.

BRAZIL. Law No. 5,966 of 11 December 1973. Establishes the National System of Metrology, Standardization and Industrial Quality, and provides for other matters. Official Gazette of the Union: Section 1, Brasília, DF, 13 December 1973.

BRAZIL. Law No. 9,933 of 20 December 1999. Provides for the powers of Conmetro and Inmetro, establishes the fee for metrological services, and provides for other matters. Official Gazette of the Union: Section 1, Brasília, DF, 21 December 1999.

BRAZIL. National Institute of Metrology, Quality and Technology – INMETRO. Ordinance No. 302 of 12 July 2021. Approves the Technical Quality Regulation and the Conformity Assessment Requirements for Toys – Consolidated. Official Gazette of the Union, Section 1, p. 30–44, 14 July 2021.

BRAZIL. National Institute of Metrology, Quality and Technology – INMETRO. Ordinance No. 200 of 29 April 2021. Approves the General Requirements for Product Certification (RGCP) – Consolidated. Official Gazette of the Union, Section 1, p. 35–50, 30 April 2021.

BRAZIL. National Institute of Metrology, Quality and Technology. Study on the regulatory models of Europe and the United States of America, and on the influence of Industry 4.0 on the modernization of Inmetro's regulation. Rio de Janeiro: Inmetro, 2020, p. 46.

CENELEC – European Committee for Electrotechnical Standardization. European Committee for Electrotechnical Standardization (CENELEC). Bruxelas: CENELEC, 2025.

CHAMUSCA, Roberta; ASSAFIN, Danielle; SAMPAIO, Cristiane; MACHADO, Raphael. "*Critérios para a certificação de produtos na regulamentação do Inmetro*". P2P e Inovação, vol. 6, n.º 1, p. 49–69, 2019.

DU BOISPÉAN, Stéphane; MUECK, Markus; GAIE, Christoph. *Introduction to the European New Legislative Framework*. In: MUECK, Markus; GAIE, Christoph (ed.). European Digital Regulations. Cham: Springer, 2025.

EBERS, Martin. *Truly Risk-based Regulation of Artificial Intelligence: How to Implement the EU's AI Act*. European Journal of Risk Regulation. Cambridge: Cambridge University Press. p. 01-20, 2024.

EUROPEAN COMMISSION. *Blue Guide on the implementation of EU product rules 2022*. Luxembourg: Publications Office of the European Union, 2022.

EUROPEAN COMMISSION. Evaluation and Fitness Check Roadmap – Ares(2020) 6353387. Brussels: European Commission, 2020.

EUROPEAN COMMISSION. *Factual summary report on the public consultation on the New Legislative Framework (NLF).* Ares(2022)2416012, Mar. 2022.

EUROPEAN COMMISSION. *Harmonised standards*. Single Market – European Commission.

EUROPEAN COMMISSION. *New Legislative Framework*. Single Market – European Commission.

EUROPEAN UNION. Decision No 768/2008/EC of the European Parliament and of the Council of 9 July 2008 on a common framework for the marketing of products and repealing Decision 93/465/EEC. Official Journal of the European Union, 13 August 2008.

EUROPEAN UNION. Directive (EU) 2024/2853 of the European Parliament and of the Council of 23 October 2024 on liability for defective products and repealing Council Directive 85/374/EEC. Official Journal of the European Union, L, 18 November 2024.

EUROPEAN UNION. Directive 2009/48/EC of the European Parliament and of the Council of 18 June 2009 on the safety of toys. Official Journal of the European Union, 30 June 2009.

EUROPEAN UNION. European Commission. Commission Implementing Decision (EU) 2023/740 of 4 April 2023 on harmonised standards for toys drafted in support of Directive 2009/48/EC of the European Parliament and of the Council. Official Journal of the European Union, 5 April 2023.

EUROPEAN UNION. European Commission. Commission Staff Working Document – Evaluation of the New Legislative Framework. SWD(2022), Corrigendum. Brussels, 16 November 2022.

EUROPEAN UNION. Regulation (EC) No 765/2008 of the European Parliament and of the Council of 9 July 2008 setting out the requirements for accreditation and market surveillance relating to the marketing of products and repealing Regulation (EEC) No 339/93. Official Journal of the European Union, 13 August 2008.

EUROPEAN UNION. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data (General Data Protection Regulation – GDPR). Official Journal of the European Union, 4 May 2016.

EUROPEAN UNION. Regulation (EU) 2017/745 of the European Parliament and of the Council of 5 April 2017 on medical devices, and repealing Directives 90/385/EEC and 93/42/EEC. Official Journal of the European Union, 5 May 2017.

EUROPEAN UNION. Regulation (EU) 2019/1020 of the European Parliament and of the Council of 20 June 2019 on market surveillance and compliance of products and amending Directive 2004/42/EC and Regulations (EC) No 765/2008 and (EU) No 305/2011. Official Journal of the European Union, 25 June 2019.

EUROPEAN UNION. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending Regulations (EC) No 300/2008, (EU) No 2018/858, (EU) 2019/2144, (EU) 2019/881 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828. Official Journal of the European Union, 12 June 2024.

EUROPEAN UNION. Treaty on the Functioning of the European Union. Lisbon, 13 December 2007.

FAZLIOGLU, Müge. *EU AI Act Compliance Matrix*. IAPP – International Association of Privacy Professionals, 2024.

GIKAY, Asress Adimi. *Risks, innovation, and adaptability in the UK's incrementalism versus the European Union's comprehensive artificial intelligence regulation*. International Journal of Law and Information Technology. Oxford: Oxford University Press. v. 32, n. 1, p. 01-25, June, 2024.

PARENTONI, Leonardo. *O direito ao esquecimento (Right to Oblivion)*. In: LUCCA, Newton De; SIMÃO FILHO, Adalberto; LIMA, Cíntia Rosa Pereira de (org.). Direito & Internet III: Marco Civil da Internet (Lei nº 12.965/2014). 1. ed. São Paulo: Quartier Latin, 2015. p. 539–618.

SCASSA, Teresa. *Sovereignity and the Governance of Artificial Intelligence*. UCLA Law Review. Los Angeles: University of California School of Law. v. 71, n. 1, p. 214-229, May, 2024.

THORSTENSEN, Vera; ARIMA JUNIOR, Mauro Kiithi; MEGALE, Tiago Matsuoka. Conformity assessment in Brazil and Argentina: differences, similarities and convergence models (Avaliação da conformidade no Brasil e na Argentina: diferenças, similaridades e modelos de convergência). Boletim de Economia e Política Internacional (BEPI), no. 26, Sept. 2019 / Apr. 2020.

VOIGT, Paul; HULLEN, Nils. *The EU AI Act: Answers to Frequently Asked Questions*. New York: Springer Nature, 2024.

ISO/IEC 42001, DRAFT BILL NO. 2.338/2023 AND EUROPEAN AI ACT: A critical-comparative analysis of governance and regulation of artificial intelligence in Brazil

Caroline Alves Martins Pires Corrêa

Master of Laws from Federal University of Minas Gerais. Postgraduate degree in Procedural Law from the Pontifical Catholic University of Minas Gerais and in Digital Law and Data Protection from the Americas Union University Center. Researcher in the Law and Technology group at the DTEC DTIBR Center for Studies. Corporate lawyer. Information Security Analyst at NetLex.Brazil.

Lucas Carvalho Corrêa

Graduated in Information Systems from the Pontifical Catholic University of Minas Gerais. Senior Software Engineer and Architecture at LWSA Group. Brazil.

DOI: https://doi.org/10.59224/dti8.ch10

Abstract: This article analyzes the convergence between Draft Bill No. 2.338/2023, the European AI Act, and the ISO/IEC 42001 standard concerning artificial intelligence governance in Brazil. The research uses the deductive method and comparative analysis to investigate how the Brazilian framework aligns with the European risk-based approach. It concludes that Draft Bill 2.338/2023 seeks international equivalence, and that ISO/IEC 42001 is the essential technical instrument for operationalizing legal compliance and guaranteeing the protection of fundamental rights in generative language systems.

Keywords: ISO/IEC 42001. Draft Bill No. 2.338/2023. European AI Act.

Table of contents: 1. Introduction; 2. Technical Standard - Artificial Intelligence ISO/IEC 42001 (Soft Law); 3. Europe - AI Act (Hard Law); 4. Draft Bill No. 2.338/2023 (Hard Law); 5. Technical Comparison: ISO/IEC 42001 vs Pl 2.338/2023 and EU AI Act; 5.1 Risk Management; 5.2 Organizational Governance; 5.3. Ethical Principles; 5.4. Other Obligations And Duties; 5.5. Prohibited Systems; 5.6. Sanctions; 5.7. Public Sector; 5.8. Table Of Contents; Conclusion; References.

1. INTRODUCTION

Economic, political, and social systems have been influenced by changes brought about by artificial intelligence tools, especially generative programming. This is because such models have come to be used in critical sectors such as health, justice, public safety, and finance, influencing not only productivity but also decision-making. In this scenario, the discussion about AI governance and regulation has ceased to be merely theoretical and has become central to the preservation of fundamental rights, institutional stability and security, leading to the emergence of regulatory approaches on the subject globally, in different jurisdictions.

The regulation of artificial intelligence must be understood in light of the risk society theory, which analyzes the probabilities of potential harm due to certain technological processes created by humanity itself and which have global impacts (BECK, 1992, p. 4). It is within this context that artificial intelligence is situated, especially generative AI models that operate with scalability and the potential for massive global impact.

In 2024, the European Union approved the AI Act, adopting a risk-based model with a strong emphasis on protecting fundamental rights and controlling high-risk systems. Simultaneously, the International Organization for Standardization published ISO/IEC 42001:2024 in the same year, establishing a management system for artificial intelligence (AI Management System – AIMS), inspired by models such as ISO 27001. In Brazil, Draft Bill No. 2.338/2023 proposes a regulatory framework for AI, also based on risk level classification and inspired, in part, by the European experience (BRAZIL, 2025).

This study seeks to address this following question: to what extent is the legal framework proposed by Draft Bill No. 2.338/2023 aligned with the technical requirements of ISO/IEC 42001 and with the regulatory approach of the European AI Act, particularly with respect to compliance-related gaps.

The main objective of this study is to verify whether the Brazilian draft law is compatible with the technical management-system standard required in ISO/IEC 42001 and the substantive obligations of the European regulation. Secondary objectives include identifying discrepancies or normative gaps between these three instruments; verifying whether Brazil will meet international regulatory standards on artificial intelligence and potentially using this scientific work to support companies in implementing AI systems and governments in improving legislation.

The initial hypothesis is that Draft Bill No. 2.338/2023 presents partial convergences with ISO/IEC 42001 and the European AI Act, especially regard to risk management, transparency, and accountability. However, gaps remain concerning technical standardization and the granularity of risk categorization found at the AI Act, which will be analyzed throughout this work, and may compromise international harmonization and practical effectiveness of AI regulation in Brazil.

To test this hypothesis, the article proceeds in five parts. Section 2 examines ISO/IEC 42001 as an instance of soft law. Section 3 analyzes the European AI Act as a hard-law, risk-based regime. Section 4 examines Draft Bill No. 2.338/2023 and its main structural choices. Section 5 develops a technical, article-by-article comparison across five axes, like risk management, organizational governance, ethical principles, prohibited systems, and sanctions and institutional design, before the conclusion synthesizes the findings and proposes a hybrid regulatory model for Brazil.

2. TECHNICAL STANDARD - ARTIFICIAL INTELLIGENCE ISO/IEC 42001 (SOFT LAW)

Inspired by the teachings of Norberto Bobbio (BOBBIO, 1999), it is known that society is regulated by various norms, whether moral, social, religious, or legal. Within the legal category itself, other classifications and further distinctions can be drawn, but this article focuses on the classic dichotomy between soft law and hard law. Soft law refers to principles, guidelines, codes of conduct, resolutions, certifica-

tions, and audit requirements developed by public or private institutions, which lack direct state enforceability but nonetheless generate normative influence through market adoption (ARBEL, 2024, p. 609).

As already said, this work concentrates on the technical standards issued by the International Organization for Standardization (ISO), a non-governmental organization composed of representatives from national standardization bodies. The ISO standards include *AI* Management Systems (ISO/IEC 42001), AI Governance (ISO 38507) and AI Risk Management (ISO 23894). However, this work will concentrate mainly on ISO/IEC 42001 standards, which sets out requirements and guidance for establishing, implementing, maintaining and continually improving an AI management system within organizations of any size and sector.

The standard does not regulate algorithms directly, instead, it mandates the creation of organizational control processes, particularly through its Annex A. This includes controls for impact assessment, risk management, continuous monitoring, documentation governance and internal audit procedures (ISO, 2023). This approach in ISO resembles a model of "compliance by design", which means that organizations are encouraged to implement structured AI management process in pursuit of regulatory and market compliance. However, the central limitation is that a soft law is a private standard without enforceability, it does not have a coercive force, in consequence market-level adoption of ISO/IEC 42001 depends on voluntary uptake, even though it can function as an effective vehicle for good business practice. To clarify, even where compatibility with ISO/IEC 42001 is not legally mandatory, an organization may require compliance with this standard in its internal audits, as well as demand it from its suppliers and partners.

The practical effect in this described scenario ends up being, consequently, international standardization for companies at the *soft law level* is driven by market and contractual pressure rather than by state command. The following chapters examines whether the same governance requirements proposed by ISO/IEC 42001 are mirrored in the European AI Act and the Brazilian project, so that such standardization can also be extended to the regulatory level, in terms of *hard law*.

3. EUROPE - AI ACT (HARD LAW)

While ISO/IEC 42001 operationalizes ethical governance through organizational requirements structured at the level of soft law, the European AI Act transforms many of the same principles into binding legal obligations that is, into *hard law*.

Treaties, conventions, statutes, and other binding normative instruments fall under this category: they typically embody express obligations, rights, or requirements that can be monitored, recorded, and audited, thereby incorporating enforceable, subjective legal positions and reflecting the specificities of each domestic legal order responsible for their implementation highlighting the specificities of each country for the incorporation of such documents into its internal legal system (ARBEL; TOKSON; LIN, 2024, p. 613).

That said, although the European AI Act may initially appear to be a national regulatory document limited to the countries of the European Union, its impacts can be seen at a transnational level, because its effects extend transnationally (EU, 2024). [1] The strategy adopted to allow this adherence to supranationality (extraterritorial reach) while avoiding direct conflicts with the internal laws, domestic law of third countries is the adoption of a robust, principle-based structure combined with a market-access logic (VOIGT and HULLEN, 2024, p.19).

The European AI regulation applies to all suppliers from outside the European Union (EU) that place AI systems on the market or make them available in services within the EU, as per Article 2 of the regulation (EU, 2024). The EU AI Act also applies to suppliers and implementers of artificial intelligence systems established outside the European Union, provided that the output generated by these systems is used within the territory of the European Union. In the case of high-risk systems from suppliers not established in the EU, they must appoint an authorized representative in the EU before making such a system available, as per Article 54 of the (EU, 2024). Generally, high-risk systems are those that pose a threat to health, safety,

[1] Article 2: 1. This Regulation applies to: (c) providers and deployers of AI systems that have their place of establishment or are located in a third country, where the output produced by the AI system is used in the Union;(EU, 2024)

fundamental rights, or influence decision-making process as established in article 6 of the regulation (EU, 2024).

The regulatory architecture of the AI Act is best understood as a pyramid of four risk tiers, each triggering a distinct intensity of legal obligation. This hierarchical, risk-based approach rather than a single, uniform set of rules applicable to every AI system is the Regulation's central methodological choice, and it is worth describing each tier separately before turning to the specific technical requirements imposed on the highest tier, which risks are classified as unacceptable, high, limited and minimal.

The unacceptable risk is defined as prohibited because it is considered incompatible with the Union's fundamental values and doesn't have a risk mitigation plan related. Article 5 of the Regulation lists exhaustively these practices, exemplifying AI systems that have manipulative techniques capable of causing harm, exploring vulnerabilities related to age, disability, or socio-economic situation, social scoring by public authorities and real-time remote biometric identification in publicly accessible spaces are considered unacceptable (EU, 2024).

The high risk covers systems identified in Article 6 and Annex III of the Regulation. It's normally related to AI systems used in critical infrastructure, education, employment, essential private and public services, law enforcement, migration, and the administration of justice. Normally one malfunction or misuse in that scenario could affect health, safety, or fundamental rights, so it needs a strictly regulated environment with *ex ante* risk-management system.

This *ex ante* system claims for an iterative process running throughout the system's entire lifecycle (Article 9), coupled with data-governance requirements ensuring the quality and representativeness of training, validation, and testing datasets (Article 10). Providers must produce and keep up to date detailed technical documentation demonstrating compliance before the system is placed on the market (Article 11 and Annex IV), maintain automatic event logs enabling traceability (Article 12), ensure sufficient transparency for deployers to interpret the system's output and use it appropriately (Article 13), design the system to allow effective human oversight (Article 14), and guarantee an adequate level of accuracy, robustness, and cybersecurity (Article 15). To help in the implementation's practical terms the NIST AI Risk Management Sytem framework (NIST, 2023) could help to struc-

ture voluntary practices to this lifecycle formalized in a internal conformity assessment (BUTT *et all*, 2026).

There are also systems that do not meet the high-risk threshold but that pose a specific risk of deception or confusion to natural people are subject to targeted transparency duties rather than substantive risk-management obligations. Those are the systems who has limited risk, and for those cases article 50 of the Ai Act just requires, for example, that providers ensure that natural persons are informed when they are interacting with an AI system such as a chatbot, that AI-generated or manipulated audio, image, video, or text content be disclosed as artificially generated or manipulated, and that emotion-recognition and biometric-categorization systems be flagged to the individuals exposed to them. Compliance here is related to a mandatory transparency.

Finally, the minimal risk where we can found spam filters, AI-enabled video games, or inventory-management tools which providers are not subject to mandatory obligations under the AI Act, although Article 95 encourages the voluntary adoption of codes of conduct, which is precisely the space in which management-system standards such as ISO/IEC 42001 can play a meaningful role even absent legal compulsion.

The organization must have a conformity assessment, made by its internal committee, internal audit process or notified body. Crucially, high-risk systems and their providers that must be registered in the EU database for high-risk AI systems, as mentioned in Article 71 (EU, 2024). It is very clear that governance and risk management must be embedded throughout the system's lifecycle, alongside exhaustive technical documentation, as reflected in ISO/IEC 42001. Furthermore, mandatory EU-level registration is required for high-risk systems, giving this tier its distinctively hard-law and auditable.

The AI Act combines elements of public law, such as prohibitions, oversight duties, and sanctions, with technical requirements inspired by ISO-type management standards, such as risk management systems, technical documentation, record-keeping, transparency, human oversight, and accuracy/robustness/cybersecurity requirements for high-risk systems (Articles 9 to 15). And because it is a Unlike ISO/IEC 42001, this act is a *hard law regulation backed by the state enforcement*,

including significant administrative fines that can reach the higher of a fixed amount or a percentage of global annual turnover. Its model thus integrates the protection of fundamental rights with technical and operational control mechanisms enforceable by national market-surveillance authorities and the European AI Office as described in Article 74 (EU, 2024).

4. DRAFT BILL NO. 2.338/2023 (HARD LAW)

Regardless of the specific regulatory technique adopted, the pursuit of principles such as security, fairness, and transparency in the use of artificial intelligence is virtually unanimous across jurisdictions (ARBEL; TOKSON; LIN, 2024, p. 613). Draft Bill No. 2.338/2023 fits squarely within this trend, positioning itself between the soft-law logic of technical standardization and the *hard law* logic of the AI Act, because it seeks to institutionalize substantive principles and risk-based obligations, while still facing significant challenges regarding the practical implementation of conformity-assessment mechanisms.

The Brazilian Draft Bill was inspirated by the EU AI Act (FRAZÃO, 2024, p. 55) and for this reason has some points in common with that.

For example, both adopt classification by risk levels but with different levels. The EU AI Act has classified the risks into four categories like unacceptable (article 5), high-risk (article 6), minimum risk and none (EU, 2024), while the Brazilian draft distinguish between excessive-risk (prohibited) systems and high-risk systems subject to enhanced obligations (Articles 13, 14, and 18), as already explained in the number 3 above (BRAZIL, 2025).

Duties imposed on developers to prevent misuse of excessive-risk systems (Article 13, §1º) and duties to guarantee system security and the rights of affected persons (Article 17) are not identical in the EU AI ACT. Instead, the EU text splits these requirements for preventing misuse, as preparing a risk management systems and tests as told in Articles 9 and human oversight expressed in article 14, for system security (Article 15) and for provider obligations, quality management, and fundamental rights protections (Articles 16, 17 and 27). This is an important highlight when we deeply think about the cultural differences also have implications on the law, while the European document refers the security to the providers in general, the

Brazilian one address it directly to the developers.

The Draft Bill also articulates a set of substantive rights for individuals and groups affected by AI systems, set out in Article 5 the right to prior information about interactions with AI systems, the right to privacy and personal-data protection, in articulation with the Brazilian Data Protection Law No. 13.709/2018 (BRAZIL, 2018) and the right to non-discrimination and to the correction of unlawful or abusive discriminatory bias, whether direct or indirect. These rights are complemented by the duty to use standardized icons and symbols to convey the right to information (Article 5, §1º) and by heightened transparency and plain-language duties toward vulnerable groups (Article 5, §2º) (BRAZIL, 2018). This structure closely parallels the transparency and fundamental-rights protections embedded in the AI Act, although it is formulated in more open-textured, principle-based language (EU, 2024).

However, its regulatory density remains lower than that of the AI Act, particularly with respect to the technical specification of obligations, conformity assessment procedures and the quantification of sanctions, gaps that are examined in detail in Section 5.

5. TECHNICAL COMPARISON: ISO/IEC 42001 VS PL 2.338/2023 AND EU AI ACT

Having examined each instrument individually, this section undertakes a structured, article-by-article comparison across five axes that are common to all three documents minds: risk management, organizational governance, ethical principles, prohibited systems, and sanctions and institutional design. This is important because one of the most critical points mentioned in the first version of the project from 2023 concerned about it was trying to regulate AI by "not regulating", in other words, those initiatives were dramatically generical and doesn't stablished or addressed specific duties, but it focused on matters of principle which didn't addressed the core and real AI problems (FRAZÃO, 2024, p.55). So, for that case the purpose of this comparison is not merely descriptive, it seeks to identify, for each axis, whether Draft Bill No. 2.338/2023 converges with the European model, presents an operational gap that could be filled by voluntary adoption of ISO/IEC 42001, or reveals a structural deficiency that only future legislative or regulatory action can resolve as proposed in

this article.

5.1 RISK MANAGEMENT

The ISO/IEC requires organizations to establish a formal, documented risk-management process, encompassing risk identification, analysis, evaluation, and treatment, integrated into the organization's overall management system (ISO, 2023). While the Brazilian project Draft Bill No. 2.338/2023, by contract, mentions risk assessment but lacks methodological detail. Explaining a little bit more, it refers to risk assessment obligations (Articles 12 and 18)[2] without prescribing the same level of methodological detail. This is rational and acceptable when we think that ISO standards have different proposals that enforced laws, as we saw in the chapters above the differences between soft law and hard law respectively (BRAZIL, 2025). By the way, he Brazilian Draft Bill neither specifies a standardized risk-assessment methodology nor establishes mandatory documentation templates comparable to the technical documentation required under Annex IV of the EU AI Act (EU, 2024) [3] so the European document is more clear and designated to a risk assessment step by step.

There is, therefore, conceptual alignment between the three instruments, like all adopt a risk-tiered logic, but an acceptable operational gap subsists in the Brazilian text, one that could be substantially narrowed through the voluntary, complementary adoption of ISO/IEC 42001 by companies operating in Brazil. The practical consequences of this gap are not merely theoretical, as said, divergent methodological requirements across jurisdictions increase compliance costs for companies operating transnationally, which were excessively discussed by Anu Bradford and described as "the Brussels effect" (BRADFORD, 2019)[4]. And create potential non-tariff trade

[2] The article 12 tells about a preliminary assessment to address the risk before an AI product is introduced into the market. The 18, especially the item II address governance measures while the development (BRAZIL, 2025).

[3] This Annex IV is based on a legal requirement established in article 11 (EU, 2024).

[4] This is an interesting work on how extraterritorial EU regulations shape the global market. Bradford directly discusses how diverging standards force companies to choose between duplicating compliance costs across jurisdictions or adopting the strictest standard globally to reduce operat-

barriers with a "AI Productivity Maks" as highlighted by the OECD:[5]

In this scenario, to overcome this modern productivity paradox and reduce transitional friction ISO/IEC 42001 can function as a harmonizing bridge, providing a common technical vocabulary and process architecture that reduces friction between differing legal risk-classification systems and fosters greater international regulatory cooperation. The innovative environment needs its space but is doesn't has to be a trade off risk a risk assessment and precaution to mitigate, address or accept those risks, and reduce one to another is quite dangerous, irrational or even mistaken (FRAZÃO, 2024, p.68)

5.2 ORGANIZATIONAL GOVERNANCE

The ISO/IEC 42001 places strong emphasis on leadership, clearly defined roles, and organizational responsibilities as preconditions for an effective AI management system (Clause 5 of the standard) (ISO, 2023).[6] The AI Act mirrors this emphasis for high-risk systems, imposing distinct obligations on providers, deployers, importers, and distributors (Articles 26) (EU, 2023). Draft Bill No. 2.338/2023 follows a broadly similar allocation of responsibility logic, distinguishing between the developer and other "AI agents" involved in the system's lifecycle, although the precise contours of each role remain less detailed than in the European text and warrant further legislative clarification as the Draft Bill advances (article 17) (BRAZIL, 2025).

This governance dimension also encompasses incident-reporting duties. Article

ing costs.

5 "However, empirical evidence on the links between AI and productivity is only just emerging. While there is a deepening and broadening use of AI technologies in specific companies and sectors, especially data-intensive sectors such as finance, insurance and on-lineconsumer platforms, on aggregate, the benefits of transformative technologies like AI remain elusive in the productivity statistics. This has been referred to as the 'modern productivity paradox' " (OECD, 2022, p. 12)

6 5.1 Leadership and commitment: Top management shall demonstrate leadership and commitment with respect to the AI management system by: ensuring that the AI policy (see 5.2) and AI objectives (see 6.2) are established and are compatible with the strategic direction of the organization; (ISO, 2023)

42 of the Draft Bill requires AI agents to report serious security incidents to the competent sectoral authority (BRAZIL, 2025).[7] Operationalizing this duty according to ISO/IEC 42001 would mean embedding incident detection, classification, and escalation procedures directly into the organization's AI management system, so that the legal reporting deadline can realistically be met through a pre-existing internal control rather than through ad hoc, reactive compliance.

A similar operationalization gap arises with respect to the preliminary documentation required to determine a system's risk level. Article 12 of the Draft Bill provides that, prior to market introduction or use, the AI agent may carry out a preliminary risk assessment (BRAZIL, 2025).[8] ISO/IEC 42001's requirements for documented information (Clause 7.5) and for AI system impact assessments (Annex A, control A.5) offer a ready-made methodological scaffold for this preliminary assessment, again illustrating how the technical standard can operationalize an obligation that the Draft Bill states only in principle.

In this meantime, there is another ISO framework that could help. The ISO/IEC 38507 which guides the governing body (board) of organizations in evaluating, directing, and monitoring the use of artificial intelligence, adapting the principles of corporate IT governance to the AI context. The standard reinforces that AI governance should not be restricted to technical teams, but should involve senior management in defining roles, allocating responsibilities, and continuously overseeing systems in operation. This logic can directly assist the distinction still underdeveloped in Draft Bill No. 2.338/2023between the developer and the other "AI agents" mentioned in the legal text, insofar as ISO/IEC 38.507 provides parameters for the

---

[7] Art. 42. The AI agent shall report, within a timeframe to be established, to the sectoral authority the occurrence of a serious security incident, including when there is a risk to the life and physical integrity of persons, the interruption of critical infrastructure operations and serious damage to property or the environment, as well as serious violations of fundamental rights, the integrity of information, freedom of expression and the democratic process, in accordance with regulations. (BRAZIL, 2025)

[8] Art. 12. Before its introduction and circulation in the market, employment or use, the AI agent may carry out a preliminary assessment to determine the degree of risk of the system, based on the criteria set forth in this Chapter, in accordance with the state of the art of technological development. (BRAZIL, 2025)

organization to establish, in a documented manner, who is accountable for each obligation throughout the system's lifecycle, which is particularly useful for practically enabling both the duty to report serious incidents (Art. 42 of the Draft Bill) and the duty of prior risk assessment (Art. 12 of the Draft Bill) (BRAZIL, 2025). By positioning AI governance as a responsibility of senior leadership, rather than solely of development teams, ISO/IEC 38507 complements and reinforces the leadership and role-allocation requirements already set out in Clause 5 of ISO/IEC 42001, bringing the governance model of Draft Bill No. 2.338/2023 closer to a more robust and auditable accountability structure.

5.3. ETHICAL PRINCIPLES

In this section, one comparison between the principles expressed in the AI Act with those of the ISO needs to be addressed, especially highlighting the concern for fundamental rights and responsibilities in countries like Brazil, where serious harm consequences of AI using can affect sensible sectors of the society (FRAZÃO, 2024, p.56).

The ethical principles articulated in Article 5 of the Draft Bill[9] can be organized as follows: (a) prior information (Art. 5, I); (b) privacy and data protection (Art. 5, II);

[9] Article 5. Any person or group affected by an AI system, regardless of its risk level, has the following rights, to be exercised in the manner and under the conditions described in this Chapter: I – the right to information regarding their interactions with AI systems, provided in an accessible, free, and easily understandable manner—including information on the automated nature of the interaction—except in cases involving AI systems dedicated solely and exclusively to cybersecurity and cyber defense, as established by regulation; II – the right to privacy and personal data protection, particularly the rights of data subjects under the terms of Law No. 13.709 of August 14, 2018 (General Personal Data Protection Law), and relevant legislation; III – the right to freedom from illicit or abusive discrimination and to the correction of illegal or abusive discriminatory biases, whether direct or indirect. § 1. The information referred to in Item I of the *caput* of this article shall be provided using standardized, easily recognizable icons or symbols, without prejudice to other formats. § 2. AI systems intended for vulnerable groups must, at all stages of their lifecycle, be transparent, employ simple and clear language appropriate to the users' age and cognitive capacity, and be implemented with due regard for the best interests of such groups. (BRAZIL, 2025)

(c) non-discrimination and correction of discriminatory bias (Art. 5, III); (d) use of standardized icons and symbols to convey the right to information (Art. 5, §1º); and (e) transparency and plain language toward vulnerable groups (Art. 5, §2º) (BRAZIL, 2025).

These provisions substantially converge with the fundamental-rights orientation of the AI Act, particularly its transparency obligations for AI systems interacting with natural persons (Article 13)[10] (EU, 2023) and its emphasis on non-discrimination throughout the high-risk conformity-assessment process (Article 11 and Annex IV). ISO/IEC 42001, by contrast, expresses its own set of principles in organizational rather than fundamental-rights terms, for instance, accountability, fairness, transparency, and human oversight are framed as management-system objectives to be pursued through policy, competence, and monitoring requirements (Clauses 5 and 6 of the standard). The comparison is nonetheless instructive because it shows how the normative principles enshrined in the Draft Bill and the AI Act can be operationalized through the organizational mechanisms already codified in ISO/IEC 42001, so that abstract commitments to non-discrimination or transparency translate into concrete controls, competence requirements, and monitoring indicators inside the organization.

One of the greatest day-to-day challenges for cybersecurity professionals is translating legal requirements into measurable operational controls, as proposed at NIST AI Framework (NIST, 2023). This challenge becomes even greater when addressing respect for rights, ranging from a simple, user-configurable button that adds a symbol indicating AI-generated content, to understanding data transparency regarding model training to ensure the absence of bias.

[10] Article 13. Transparency and provision of information to deployers.1. High-risk AI systems shall be designed and developed in such a way as to ensure that their operation is sufficiently transparent to enable deployers to interpret a system's output and use it appropriately. An appropriate type and degree of transparency shall be ensured with a view to achieving compliance with the relevant obligations of the provider and deployer set out in Section 3. (EU, 2023)

5.4. OTHER OBLIGATIONS AND DUTIES

Beyond the ethical principles set out in Article 5, the Draft Bill imposes a set of additional obligations that further approximate it to the AI Act's compliance architecture: preparation of preliminary documentation to determine the system's risk level (Art. 12); prohibition on the development of excessively risky systems (Art. 13, §1º) [11]; the guarantee of system security and the fulfillment of the rights of affected individuals and groups (Art. 17)[12]; and finally, the duty to communicate serious incidents (Art. 42), insofar as it is simultaneously a governance duty and a substantive obligation toward affected persons (BRAZIL, 2025).

Here an important point is from *who* tell will need to communicate the incident. The initial version of the Draft Bill introduced in 2023 (BRAZIL, 2025) contained no express mention of this competent authority. However, the situation changed in January 2025 when the new Senate text introduces the National Data Protection Authority (ANPD), as well as direct responsibility for specific government entities, depending on the nature of the matter. That was an important change because "the regulation of AI is not incompatible with the initiatives and particularities of each economic agent" (FRAZÃO, 2024, p. 56), so the specificity of each sector should be addresses and valuated, to ensure harmonization between the innovation.

The article 42 of Draft Bill No. 2.338/2023 establishes that the AI agent must report serious security incidents to the "sectoral authority" ("autoridade setorial"), within a timeframe still to be defined by regulation. The provision covers incidents that pose a risk to the life and physical integrity of people, the interruption of critical infrastructure operations, serious damage to property or the environment, and serious violations of fundamental rights, the integrity of information, freedom of expression, and the democratic process. Read in isolation, however, Article 42 does not identify, in a first moment, which body ultimately coordinates this reporting duty

[11] Art. 13 (...) § 1º AI system developers must adopt measures to prevent the use of their systems for the situations described in the heading of this article. (BRAZIL, 2025)

[12] Art. 17. AI agents must guarantee the security of systems and the protection of the rights of affected individuals or groups, in accordance with regulations. (BRAZIL, 2025)

at the national level, a gap that can only be resolved by reading it together with the Draft Bill's institutional architecture set out in Chapter IX, which could be remedied by an express, direct, and clear mention (BRAZIL, 2025).

It's important to mention that articles 45 to 47 supply this missing layer, specially the article 45, §1º, I that designates the National Data Protection Authority (ANPD) as the "competent authority" ("autoridade competente") responsible for coordinating the National System for the Regulation and Governance of Artificial Intelligence (Sistema Nacional de Regulação e Governança de Inteligência Artificial SIA), which also integrates the sectoral authorities, the Permanent Council for AI Regulatory Cooperation (CRIA), and the Committee of AI Experts and Scientists (CECIA).

On the same side, article 46, II, c, further empowers the ANPD to issue binding general rules on "procedures for the communication of serious incidents" (BRAZIL, 2025), meaning that, while the day-to-day recipient of a given incident report is ordinarily the sectoral authority governing that specific market, it is the ANPD that standardizes and harmonizes how that reporting duty operates nationwide. Furthermore, article 47 [13]completes the architecture by attributing to the ANPD a residual regulatory role, which means whenever a given economic activity lacks a specific sectoral regulator, full normative, regulatory, supervisory, and sanctioning competence including, therefore, the direct receipt of incident reports, falls to the ANPD itself (BRAZIL, 2025).

This two layer design bears a structural resemblance to the reporting architecture of the EU AI Act, while also departing from it in a meaningful way. Under article 73 of the European document[14], providers of high-risk systems report serious incidents directly to the market surveillance authority of the Member State concerned, while deployers who detect such an incident must notify both the provider and the competent market surveillance authority without undue delay, as mentioned in the article

[13] Art. 47. In the capacity of residual regulator, the competent authority shall exercise full norm-setting, regulatory, supervisory, and sanctioning powers regarding the development, implementation, and use of AI systems for economic activities where there is no specific sectoral regulatory body or entity. (BRAZIL, 2025)

[14] Article 73. Reporting of serious incidents: 1. Providers of high-risk AI systems placed on the Union market shall report any serious incident to the market surveillance authorities of the Member States where that incident occurred. (EU, 2024)

26 (5)[15] (EU, 2024). The AI Act's coordination at Union level is entrusted to the European AI Office, but no single national authority is designated as the default or residual recipient of incident reports across all economic sectors within a given Member State.

At the brazilian side, the Draft Bill No. 2.338/2023, by decentralizing the primary reporting duty to sectoral authorities while vesting the ANPD with both a harmonizing, rule-making function and a residual, catch-all competence, produces a more centralized and predictable institutional reference point than the AI Act's Member-State-by-Member-State model a distinctive feature of the Brazilian proposal that merits recognition in any comparative assessment of the two regimes' governance architecture. This Brazilian framework allows for the necessary flexibility regarding innovation; thus, the bill establishes a regulatory regime based on risks and rights to ensure legal certainty, while leaving room for the inevitable updates that will occur in a constantly evolving technological world (FRAZÃO, 2022, p. 69).

Taking together, these obligations reveal a Draft Bill that is conceptually complete it addresses documentation, prevention, security, and incident communication but that leaves the operational "how" of compliance largely to future regulation. This is precisely the space that ISO/IEC 42001 is designed to occupy, its Annex A controls for resource allocation, data quality, third-party and customer relationships, and incident management provide concrete, auditable content for duties that the Draft Bill currently states only at the level of principle.

[15] Article 26 - 5. Deployers shall monitor the operation of the high-risk AI system on the basis of the instructions for use and, where relevant, inform providers in accordance with Article 72. Where deployers have reason to consider that the use of the high-risk AI system in accordance with the instructions may result in that AI system presenting a risk within the meaning of Article 79(1), they shall, without undue delay, inform the provider or distributor and the relevant market surveillance authority, and shall suspend the use of that system. Where deployers have identified a serious incident, they shall also immediately inform first the provider, and then the importer or distributor and the relevant market surveillance authorities of that incident. If the deployer is not able to reach the provider, Article 73 shall apply mutatis mutandis. This obligation shall not cover sensitive operational data of deployers of AI systems which are law enforcement authorities. (EU, 2024)

5.5. PROHIBITED SYSTEMS

The AI Act explicitly lists prohibited practices in Article 5, ranging from subliminal manipulation and exploitation of vulnerabilities to social scoring by public authorities and, subject to narrow exceptions, real-time remote biometric identification in publicly accessible spaces. Article 13 of the Draft Bill[16], adopts a strikingly similar list of prohibited or "excessive-risk" systems: manipulation and exploitation of vulnerabilities (I, a-b); criminal-recidivism profiling (I, c); production or facilitation of child sexual abuse material (I, d); illegitimate social scoring by public authorities (II); autonomous weapons systems (III); and real-time remote biometric identification in public spaces, subject to exceptions comparable to those found in the European text (IV).

This near-identity of prohibited categories is one of the clearest indicators of Draft

[16] Article 13. The development, implementation, and use of AI systems are prohibited: I – for the purpose of: a) instigating or inducing the behavior of a natural person or groups in a way that causes harm to their health, safety, or other fundamental rights, or those of third parties; b) exploiting any vulnerabilities of a natural person or groups with the objective or effect of inducing their behavior in a way that causes harm to their health, safety, or other fundamental rights, or those of third parties; c) evaluating the personality traits, characteristics, or past behavior, criminal or not, of individuals or groups, for the purpose of assessing the risk of committing crimes, offenses, or recidivism; d) enabling the production or dissemination, or facilitating the creation, of material that characterizes or represents the sexual abuse or exploitation of children and adolescents; II - by public authorities, to evaluate, classify or rank natural persons, based on their social behavior or attributes of their personality, through universal scoring, for access to goods and services and public policies, in an illegitimate or disproportionate manner; III - in autonomous weapons systems (AWS); IV - in remote, real-time biometric identification systems in spaces accessible to the public, except in the following cases: a) instruction of an inquiry or criminal proceeding, with prior and reasoned judicial authorization, when there are reasonable indications of authorship or participation in a criminal offense, the evidence cannot be obtained by other available means and the investigated fact does not constitute a minor offense; b) search for victims of crimes and missing persons, or in circumstances involving a serious and imminent threat to the life or physical integrity of natural persons; c) flagrant offense of crimes punishable by imprisonment for more than 2 (two) years, with immediate notification to the judicial authority; d) recapture of escaped defendants and execution of arrest warrants and restrictive measures ordered by the Judiciary. (BRAZIL, 2025)

Bill's deliberate alignment with the European risk-based paradigm. ISO/IEC 42001 does not, and by design cannot, establish prohibited use-cases of this kind, since it is a management-system standard rather than a substantive rights instrument; this axis of comparison therefore confirms that the prohibition of unacceptable-risk systems is a matter properly reserved to hard law.

5.6. SANCTIONS

The EU AI Act establishes a detailed and quantified sanctioning regime, with administrative fines calibrated according to the gravity of the infringement and the size of the offending undertaking, reaching, for the most serious violations, the higher of a fixed ceiling or a percentage of global annual turnover (Articles 99 until 100). The Draft Bill No. 2.338/2023 establish an equivalent quantified sanctioning framework in article 50, not to mention civil liability existing.

Table 1: Comparative Matrix of Obligations: Draft Bill No. 2.338/2023 vs. EU AI Act vs. ISO (38507 and 42001)

Comparison Aspect	EU AI Act (European Union)	PL 2338/2023 (Brazil)
Main Articles on Penalties	Articles 99 and 101	Article 50 in the Substitute Bill for approval
Maximum Monetary Penalty Limit	€35,000,000 or 7% of total annual revenue (whichever is greater).	R$ 50,000,000 per infraction or 2% of revenue in Brazil.
Grading of Fines Based on Severity	• Prohibited Practices: Up to €35 million / 7% of revenue. • High-Risk AI Violation / Obligations: Up to €15 million / 3% of revenue. • False/Misleading Information: Up to €7.5 million / 1% of revenue.	The legal text does not explicitly specify a scaling of infractions; the gradation of the value (up to a ceiling of R$ 50 million / 2%) is defined in the administrative assessment.
Treatment for SMEs / Startups	The lower of the fixed ceiling in euros	Parameter considered in sentencing criteria (economic condition of the

Comparison Aspect	EU AI Act (European Union)	PL 2338/2023 (Brazil)
	and the percentage of turnover applies.	offender).
Sanctions for General Use Models (GPAI)	Article 101: Fines of up to €15 million or 3% of global turnover imposed by the EU AI Office.	Included in the general regime of Article 50 (applicable to all AI agents).
Non-Monetary Sanctions	Warnings, restrictions, and market closure/retreat for non-compliant systems.	Warning, daily fine, public disclosure of the infraction, suspension/prohibition of system operation, prohibition of data processing, and prohibition/restriction from participating in the regulatory sandbox regime (up to 5 years)
Articles on Dosimetry / Criteria	Article 99(7): Takes into account gravity, duration, intentionality, cooperation and degree of damage.	Art. 50, §1 in the Substitute Bill: Assesses severity, good faith, recidivism, advantage gained, mitigation, and prior governance.
Civil Liability and Damages	Addressed in separate directives within the European bloc (AI Liability Directive).	Articles 35 to 39: Define strict liability (regardless of fault) for high-risk AI and presumed fault for other cases.

This is arguably the axis of comparison is most pronounced, since the deterrent effect of any risk-based regime depends substantially on the credibility and proportionality of its sanctions. ISO/IEC 42001 offers no substitute for this gap either, as certification under the standard entails, at most, the loss of certification status or contractual consequences, not state-imposed administrative penalties.

5.7. PUBLIC SECTOR

Analyze how the creation of regulatory authority works as expressed in the draft law. This is a Brazilian peculiarity that in itself has significant implications that would warrant a new article. However, it is interesting to mention because the AI Act does not address specific obligations for the use of AI in the public sector.

A further peculiarity of the Brazilian Draft Bill is its institutional design for enforcement. The text envisions the creation and designation of a competent regulatory authority responsible for overseeing compliance, coordinating with sectoral regulators, and issuing complementary regulation on risk assessment, incident reporting, and sanctions. This administrative authority model has no direct equivalent in the AI Act's own text, which instead relies on a composite structure of national market-surveillance authorities coordinated by a European-level AI Office.

Equally distinctive is the fact that the AI Act does not establish a dedicated, free-standing chapter of specific obligations for the use of AI by public administration bodies beyond the general high-risk classification criteria that already capture many public-sector use-cases, for example, systems used in law enforcement, migration, or the administration of justice, per Annex III. Draft Bill No. 2.338/2023, in the Brazilian constitutional and administrative-law tradition, pays closer attention to the specific accountability of public agents deploying AI systems. This is a feature that, on its own, would merit a dedicated study, for the purposes of this article, it suffices to note that the emphasis on public-sector accountability is a genuinely Brazilian contribution that goes beyond what strict transposition of the European model would have required.

5.8. Table of contents

Taken together, the seven axes examined above show that Draft Bill No. 2.338/2023 converges with the EU AI Act most strongly at the level of principles, risk categories, and prohibited practices, while an operational gap persists in risk-assessment methodology, conformity assessment, and the quantification of sanctions, precisely the space in which the voluntary adoption of ISO/IEC 42001 can operationalize obligations that the Draft Bill currently states only in principle. The comparison also reveals genuinely Brazilian institutional choices, such as the coordinating and residual role assigned to the ANPD, that go beyond a mere transposition of the European model. This composite picture partial convergence, technical gaps, and distinctive institutional design sets the stage for the concluding remarks that follow.

Table 2: Comparative Matrix of Obligations: Draft Bill No. 2.338/2023 vs. EU AI Act vs. ISO (38507 and 42001)

Dimension	Draft Bill No. 2.338/2023 (Brazil)	EU AI Act – Regulation (EU) 2024/1689	ISO/IEC 38507 and ISO/IEC 42001
Legal nature	Statute (hard law), still in the legislative process; enforcement architecture largely dependent on future regulation.	Directly applicable Regulation (hard law) already in force; self-executing across all Member States (EUROPEAN UNION, 2024).	Voluntary technical standards (soft law); certifiable management-system framework (ISO/IEC 42001) and board-level governance guidance (ISO/IEC 38507), with no binding legal force absent contractual or market adoption.
Territorial / extraterritorial scope	Applies to the development, implementation, and use of AI systems within Brazilian territory; extraterritorial criteria not yet detailed with the same precision as the European text.	Applies to providers and deployers established outside the EU whenever the system is placed on the EU market or its output is used within the Union (Art. 2); high-risk providers outside the EU must appoint an EU-based authorized representative (Art. 54).	Globally applicable regardless of jurisdiction; scope is defined by the organization's own context (ISO/IEC 42001, Clause 4), not by territory, and adoption depends on the organization's decision to implement or certify an AI management system.
Risk classification method	Risk-tiered structure distinguishing excessive-risk (prohibited) and high-risk systems (Arts. 13, 14, 18); risk-assessment methodology stated in principle, without a standardized technical procedure.	Four-tier pyramid — unacceptable, high, limited, and minimal risk — with detailed classification criteria and an exhaustive list of high-risk use-cases (Art. 6 and Annex III).	ISO/IEC 42001 requires a documented, continuous risk-assessment and treatment process (Clause 6.1) but does not itself define fixed risk tiers; ISO/IEC 38507 instructs the governing body to calibrate the intensity of oversight to the level of risk involved, without prescribing legal thresholds.
Prohibited / unacceptable-risk	Article 13 prohibits manipulation and exploitation of vulnerabilities, criminal-	Article 5 prohibits a materially similar list: subliminal/manipulative tech-	Neither standard establishes prohibited use-cases, since this is a matter reserved to

Dimension	Draft Bill No. 2.338/2023 (Brazil)	EU AI Act – Regulation (EU) 2024/1689	ISO/IEC 38507 and ISO/IEC 42001
systems	recidivism profiling, CSAM-related systems, illegitimate social scoring by public authorities, autonomous weapons systems, and (subject to exceptions) real-time remote biometric identification in public spaces.	niques, exploitation of vulnerabilities, social scoring by public authorities, and, subject to narrow law-enforcement exceptions, real-time remote biometric identification in publicly accessible spaces.	hard law; ISO/IEC 38507 directs the governing body to identify unacceptable or high-risk uses and escalate them for a strategic decision, but does not itself impose a legal ban.
High-risk obligations	General duties to prevent misuse (Art. 13, §1º) and to guarantee system security and the rights of affected persons (Art. 17); preliminary risk documentation encouraged but not mandatorily detailed (Art. 12).	Structured risk-management system throughout the system's lifecycle (Art. 9), data governance (Art. 10), technical documentation (Art. 11, Annex IV), record-keeping/logging (Art. 12), transparency to deployers (Art. 13), human oversight (Art. 14), and accuracy/robustness/cybersecurity (Art. 15).	ISO/IEC 42001 Annex A operationalizes an equivalent package through organizational controls on AI policies, resources and competence, impact assessment, the AI system life cycle, data for AI systems, and information to interested parties; ISO/IEC 38507 requires the governing body to evaluate, direct, and monitor these controls at the strategic level.
Conformity assessment and registration	Conformity-assessment procedure or public registry yet established in the Draft Bill's text;	Mandatory conformity assessment (internal or via notified body) before the CE marking is affixed, and mandatory registration of high-risk systems in the EU database (Art. 71).	ISO/IEC 42001 enables voluntary third-party certification of the AI management system by accredited certification bodies, verified through internal audit and management review (Clauses 9 and 10), but does not require public registration; certification remains market-driven rather than state-mandated.
Transpar-	General right to prior infor-	Dedicated transparency tier:	ISO/IEC 42001 Annex A

Dimension	Draft Bill No. 2.338/2023 (Brazil)	EU AI Act – Regulation (EU) 2024/1689	ISO/IEC 38507 and ISO/IEC 42001
ency duties (limited-risk systems)	mation about AI interactions (Art. 5, I) and duty to use standardized icons/symbols (Art. 5, §1º), applicable across risk levels rather than as a distinct tier.	disclosure of AI interaction (chatbots), labelling of AI-generated/manipulated content (deepfakes), and disclosure of emotion-recognition/biometric-categorization systems (Art. 50).	requires the organization to provide adequate information to interested parties about AI system operation; ISO/IEC 38507 reinforces transparency and accountability of governance decisions to stakeholders, though neither standard mandates specific disclosure formats comparable to Art. 50 of the AI Act.
Rights of affected persons	Article 5: prior information, privacy/data protection (in articulation with the LGPD), and non-discrimination/correction of discriminatory bias.	Rights embedded across the Regulation: transparency (Art. 13, Art. 50), human oversight (Art. 14), and non-discrimination pursued through data-governance and quality requirements for high-risk systems (Art. 10).	Neither ISO standard creates individual rights directly, as both are organizational governance/management instruments; ISO/IEC 42001's impact-assessment control and ISO/IEC 38507's duty to consider stakeholder and societal impact nonetheless supply the internal process through which such rights can be operationalized.
Incident reporting	Mandatory reporting of serious security incidents to the competent sectoral authority, timeframe to be further regulated (Art. 42), with the ANPD coordinating procedures nationally and acting as residual authority (Arts. 45-47).	Mandatory reporting of serious incidents involving high-risk systems to market-surveillance authorities, within defined deadlines (Art. 73).	ISO/IEC 42001 requires internal nonconformity, corrective-action, and performance-evaluation processes (Clauses 9 and 10) that can feed statutory reporting deadlines, but does not itself impose an external reporting duty to any regulator; ISO/IEC 38507 assigns to the

Dimension	Draft Bill No. 2.338/2023 (Brazil)	EU AI Act – Regulation (EU) 2024/1689	ISO/IEC 38507 and ISO/IEC 42001
			governing body accountability for ensuring such internal processes exist.
Sanctions	The legal text does not explicitly specify a scaling of infractions; the gradation of the value (up to a ceiling of R$ 50 million / 2%) is defined in the administrative assessment.	Quantified administrative fines, up to the higher of a fixed ceiling or a percentage of global annual turnover, calibrated by the gravity of the infringement (Arts. 99–100).	Neither standard entails state-imposed sanctions; the only consequence under ISO/IEC 42001 is suspension or withdrawal of certification by the certifying body, with associated reputational or contractual effects, while ISO/IEC 38507 has no sanctioning mechanism of its own.
Institutional design / enforcement authority	Envisions a competent national regulatory authority (ANPD) coordinating a broader national system (SIA) together with sectoral authorities, with a distinctive emphasis on public-sector accountability.	Decentralized model: national market-surveillance authorities in each Member State coordinated by a European-level AI Office.	Enforcement relies on private, market-based verification through accredited certification bodies (ISO/IEC 42001) and on internal accountability of the organization's own governing body (ISO/IEC 38507), rather than on a public regulator a structural contrast with both the ANPD/SIA model and the AI Act's market-surveillance authorities.

Source: prepared by the authors based on BRAZIL (2025), EUROPEAN UNION (2024), and ISO (2022, 2023).

CONCLUSION

Effective AI governance requires the integration of technical standardization and state legislation. ISO/IEC 42001 provides a robust operational framework for organi-

zations, translating abstract governance principles into auditable management processes, while the AI Act establishes a coercive, rights-protective regulatory model backed by significant sanctions.

Draft Bill No. 2.338/2023 represents meaningful progress toward international regulatory equivalence: it reproduces the European risk-tiered classification, a comparable catalogue of prohibited systems, and a set of substantive rights for affected individuals. It nevertheless still requires considerable technical refinement particularly regarding risk-assessment methodology and conformity-assessment procedures, before it can be said to achieve structural equivalence with the AI Act.

The article's hypothesis is therefore confirmed only in part. The convergence is strong at the level of principles, risk categories, and prohibited practices, but an operational gap persists in matters of technical standardization, documentation, and enforcement. Consolidating a hybrid regulatory model one that combines the binding force of Draft Bill No. 2.338/2023 with the technical operationalization offered by voluntary adoption of ISO/IEC 42001 appears to be the most consistent path for Brazil to pursue international harmonization while safeguarding the fundamental rights of those affected by artificial intelligence systems.

REFERENCES

ARBEL, Yonathan; MATTHEW, Tokson; LIN, Albert. *Systemic Regulation of Artificial Intelligence. Arizona State Law Journal* . Phoenix: Arizona State University. v. 56, no. 2, p. 545-620, Summer, 2024.

BECK, Ulrich. *Risk society*: towards a new modernity. London: Sage Publications, 1992.

BOBBIO, Norberto. *Teoria do Ordenamento Jurídico*. 10. ed. Brasília: Editora Universidade de Brasília, 1999

BRADFORD, Anu. The Brussels effect: how the European Union rules the world. New York: Oxford University Press, 2019. Available at: <https://academic.oup.com/book/36491>. Accessed on: July 10, 2026.

BRAZIL. Draft Bill No. 2.338, of May 3, 2023. *Provides for the development, promotion and ethical and responsible use of artificial intelligence based on the centrality of the human person.* Federal Senate. Text forwarded to the Chamber of Deputies, 2025. Available at: < https://www.camara.leg.br/proposicoesWeb/prop_mostrarintegra?codteor=2881712&filename=Avulso+PL+2338%2F2023 >. Accessed on: March 2, 2026.

BRAZIL Draft Bill No. 13.709, of August 14, 2018. *General Personal Data Protection Law (LGPD).* Official Gazette of the Union, Brasília, Federal District, August 14, 2018. Available at: <https://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/l13709.htm>. Accessed on: March 20, 2026.

BUTT, Talal Ashraf. IQBAL, Muhammad. IQBAL, Razi. UGAF-ITS: *A Standards Harmonization Framework and Validation Tool for Multi-Framework AI Governance in Distributed Intelligent Transportation Systems*. 7 abr. 2026. Available at: <https://arxiv.org/abs/2604.22789>. Accessed on: March 20, 2026.

EUROPEAN UNION. *Regulation (EU) 2024/1689*. European Parliament and of the Council of 13 June 2024 laying down harmonized rules on artificial intelligence (Artificial Intelligence Act). Official Journal of the European Union, 2024.

FRAZÃO, Ana. Regulation of artificial intelligence in Brazil: examination of Draft Bill no. 2338/2023. UNIO. EU Law Journal, v. 10, n. 1, p. 54-69, may 2024. Available at: < https://www.researchgate.net/publication/382094171_Regulation_of_artificial_intelligence_in_Brazil_examination_of_Draft_Bill_no_23382023>. Accessed on: July 10, 2026.

INTERNATIONAL ORGANIZATION FOR STANDARDIZATION. *(ISO)/IEC 42001:2023*. Information technology. Artificial intelligence Management system. International Organization for Standardization. Geneva, 2023.

INTERNATIONAL ORGANIZATION FOR STANDARDIZATION. *(ISO)/IEC 38507:2022.* Information technology Governance of IT. Governance implications of the use of artificial intelligence by organizations. Geneva, 2022.

NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY (NIST). *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*. Department of Commerce. United Stated of America, 2023. Available at: < https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf>. Accessed on: Abril 20, 2026.

OECD. International Regulatory Co-operation and International Trade: The Case of Artificial Intelligence Governance. OECD Trade Policy Papers, OECD Publishing, Paris, 2022. Available at: <https://www.oecd.org/en/publications/artificial-intelligence-and-international-trade_13212d3e-en.html>. Accessed on: July 10, 2026.

VOIGT, Paul; HULLEN, Nils. *The EU AI Act*: Answers to Frequently Asked Questions. New York: Springer Nature, 2024.

IV
COMPARATIVE AND GLOBAL PANORAMA: ASIA AND PACIFIC

Between Fragmentation and Convergence: Australia's Role in Asia–Pacific AI Governance

Grace Li

Professor of Law, Faculty of Law, University of Technology Sydney (UTS), Australia.

DOI: https://doi.org/10.59224/dti8.ch11

Abstract: AI governance in the Asia–Pacific is developing through divergent regulatory strategies, ranging from ethics and guidance instruments to binding statutory and security-oriented controls. Within this heterogeneous landscape, Australia has retained a principles-first and sectoral model, while progressively institutionalising AI governance through national policy, regulator guidance, assurance mechanisms, and dedicated safety capability. Drawing on a structured comparison of Japan, Singapore, South Korea, New Zealand, India, and China, the chapter argues that Australia is best understood as a bridging jurisdiction. Australia's contribution lies in governance translation: converting high-level trust and safety norms into implementable routines through regulator guidance, standards-aligned practices, and procurement and assurance mechanisms, rather than exporting a single legislative model. The chapter's practical contribution is an operational account of how Australia's soft-law frameworks interact with privacy, consumer protection, competition, online safety, anti-discrimination, and administrative-law accountability to produce de facto AI regulation. It concludes by identifying conditions under which Australia can support regional interoperability and soft convergence while maintaining domestic legitimacy and public trust.

Keywords: artificial intelligence governance; Asia–Pacific; comparative regulation; soft law; middle powers; regulatory fragmentation

Table of contents: 1. Introduction: Australia and the Asia–Pacific AI Governance

Moment. 2. The Asia–Pacific AI Governance Landscape: A Spectrum of Approaches. 2.1 Principles-Based and Guidance-Oriented Models. 2.2 Hybrid Innovation–Regulation Frameworks. 2.3 Centralised and Prescriptive Regulatory Approaches. 2.4 Regional Challenges and Fragmentation. 3. Australia's AI Governance Architecture. 3.1 Sectoral Regulation as De Facto AI Law. 3.2 Strengths and Limitations of Australia's Approach. 4. Comparative Perspectives: Australia and Key Asia–Pacific Jurisdictions. 4.1 Australia and Japan: Normative Alignment, Institutional Divergence. 4.2 Australia and Singapore: Pragmatism and Operationalisation. 4.3 Australia and South Korea: Regulatory Readiness and State Capacity. 4.4 Australia and New Zealand: Shared Values, Different Regulatory Emphases. 4.5 Australia and India: Developmental Priorities and Regulatory Sequencing. 4.6 Australia and China: Contrasting Governance Logics. 4.7 Comparative Insights. 5. Regional Cooperation, Soft Convergence, and Australia as a Bridging Jurisdiction. 5.1 Soft Convergence through Multilateral Dialogue. 5.2 Standards and Functional Interoperability. 5.3 Australia's Bridging Role in Regional AI Governance. 5.4 Strategic Implications. 6. Conclusion.

1. INTRODUCTION: AUSTRALIA AND THE ASIA–PACIFIC AI GOVERNANCE MOMENT

AI has become a defining regulatory challenge of the twenty-first century, reshaping economic activity, public administration, national security, and social relations.[1] Nowhere is this challenge more complex than in the Asia–Pacific region. Home to some of the world's most technologically advanced economies alongside rapidly digitising developing states, the region encompasses a wide spectrum of political systems, regulatory capacities, and normative commitments.[2] As a result, approaches to AI governance across the Asia–Pacific vary significantly in form, ambition, and enforcement.[3]

---

[1] OECD, OECD Principles on Artificial Intelligence (2019, updated 2024) https://oecd.ai/en/ai-principles.

[2] UNESCO, *Recommendation on the Ethics of Artificial Intelligence* (2021) https://www.unesco.org/en/artificial-intelligence/recommendation-ethics.

[3] OECD, *Framework for the Classification of AI Systems* (2022) https://www.oecd.org/going-dig-

Against this backdrop, Australia presents a particularly interesting case. Unlike jurisdictions that have adopted comprehensive AI legislation or highly centralised regulatory controls, Australia has thus far resisted a single, binding AI statute.[4] Instead, it has relied on a combination of principles-based guidance, iterative policy consultation, and the application of existing legal frameworks to manage AI-related risks.[5] This approach reflects long-standing features of Australia's regulatory culture - a preference for flexible governance tools, strong reliance on independent regulators, and sensitivity to the trade-offs between innovation, rights protection, and regulatory burden.[6]

At the same time, Australia is deeply embedded in regional and multilateral governance networks. It participates actively in forums such as the Asia-Pacific Economic Cooperation (APEC), the Quadrilateral Security Dialogue (Quad), and ASEAN–Australia digital cooperation initiatives, while aligning its domestic AI policy discourse with international principles developed by the Organisation for Economic Co-operation and Development (OECD) and other global standard-setting bodies.[7] This dual orientation, domestic regulatory caution combined with outward-facing engagement, raises important questions about Australia's role in shaping the future of AI governance in the Asia–Pacific.

This chapter argues that Australia occupies a strategic middle position within the regional AI governance landscape. It neither mirrors the prescriptive, state-directed regulatory models exemplified by China nor adopts a comprehensive, cross-sector AI statute of the kind developed in the European Union. Nor is its approach purely

ital/ai/classification-of-ai-systems.htm.

4 Australian Government, Department of Industry, Science and Resources, *Australia's AI Ethics Principles* (2019) https://www.industry.gov.au/publications/australias-ai-ethics-principles.

5 Australian Government, Safe and Responsible AI in Australia – Consultation Paper (2023) https://consult.industry.gov.au/supporting-responsible-ai.

6 Australian Law Reform Commission, *Corporate Criminal Responsibility* (ALRC Report 136, 2020) https://www.alrc.gov.au/wp-content/uploads/2020/05/ALRC-CCR-Final-Report-websml.pdf.

7 APEC, *AI Governance: Why Cooperation Matters* (2024) https://www.apec.org/press/features/2024/ai-governance--why-cooperation-matters; Australian Government Department of the Prime Minister and Cabinet, Quad Leaders' Summit, *Critical and Emerging Technology* (2023) https://www.pmc.gov.au/resources/quad-leaders-summit-2023/critical-and-emerging-technology.

market-led. Instead, Australia combines principles-based guidance, existing general and sectoral law, and increasingly formalised assurance and institutional mechanisms. This combination supports the chapter's central claim that Australia can function as a bridging jurisdiction, promoting interoperability, transparency, and accountable implementation across divergent governance systems.[89]

The chapter proceeds by situating Australia within the broader spectrum of Asia–Pacific AI governance approaches. It first outlines the regional diversity of regulatory philosophies, distinguishing between principles-based models, hybrid innovation-regulation strategies, and more centralised or prescriptive regimes. It then examines Australia's AI governance architecture in detail, focusing on the interaction between soft-law instruments, such as the AI Ethics Principles and the Safe and Responsible AI framework, and sectoral regulation in areas including privacy, consumer protection, competition, online safety, and anti-discrimination law.[10] Building on this foundation, the chapter undertakes a comparative analysis of Australia's approach alongside those of Japan, Singapore, South Korea, New Zealand, India, and China, highlighting points of convergence, divergence, and mutual learning.[11]

Finally, the chapter explores the implications of Australia's position for regional cooperation and future regulatory development. It assesses whether Australia's

---

8 Cyberspace Administration of China, *Provisions on the Administration of Algorithmic Recommendation Services* 互联网信息服务算法推荐管理规定 (2022) https://www.cac.gov.cn/2022-01/04/c_1642894606364259.htm.

9 Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act), OJ L 2024/1689, 12 July 2024, https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng.

10 Office of the Australian Information Commissioner, Artificial Intelligence and Privacy Guidance https://www.oaic.gov.au/privacy/privacy-guidance-for-organisations-and-government-agencies/guidance-on-privacy-and-the-use-of-commercially-available-ai-products; ACCC, Digital Platform Services Inquiry – Final Report (2025) https://www.accc.gov.au/system/files/digital-platform-services-inquiry-final-report-march2025.pdf.

11 Government of Japan, *Social Principles of Human-Centric AI* (2019) https://www.cas.go.jp/jp/seisaku/jinkouchinou/pdf/humancentricai.pdf; New Zealand Government, *Algorithm Charter for Aotearoa New Zealand* (2020) https://www.data.govt.nz/toolkit/data-ethics/government-algorithm-transparency-and-accountability/algorithm-charter/.

current approach can sustain growing AI-related risks, and whether a shift toward more formalised regulation is likely or desirable. In doing so, the chapter contributes to broader debates on how middle-power jurisdictions can exercise normative influence and support the development of trustworthy AI governance in a fragmented global regulatory environment.

This chapter uses the concept of a bridging jurisdiction as an analytical device rather than a claim of regulatory leadership. A bridging jurisdiction performs three functions: translation of high-level principles into implementable governance routines; interoperability through standards and shared risk vocabularies; and brokerage through convening and capacity-building across divergent regimes.

To discipline the comparison, each jurisdiction is assessed against the same variables: regulatory posture (principles-based, hybrid, or prescriptive), institutional architecture (central coordination, regulators, and advisory bodies), enforcement and assurance tools (hard law, sandboxes, audits, registers, or transparency duties), and the dominant policy rationale (innovation facilitation, rights protection, security, or developmental sequencing). The aim is to identify where functional convergence is already emerging and where Australia can contribute by converting shared aspirations for trustworthy AI into practical, reviewable governance mechanisms.

2. THE ASIA–PACIFIC AI GOVERNANCE LANDSCAPE: A SPECTRUM OF APPROACHES

The Asia–Pacific region does not exhibit a single or convergent model of artificial intelligence (AI) governance. Instead, it reflects a broad spectrum of regulatory philosophies shaped by divergent political systems, economic priorities, technological capacities, and institutional traditions.[12] This diversity has intensified in recent years as governments across the region have moved from high-level ethical discourse toward more concrete governance mechanisms addressing the deployment, risk

[12] Aneja, U., Gupta, A. and John, S. (2025), *Exploring the Potential of AI to Usher in a New Governance Paradigm in Asia and the Pacific,* Digital Future Labs- United Nations Development Programme, https://www.undp.org/sites/g/files/zskgke326/files/2025-12/exploring-the-potential-of-ai-to-usher-in-a-new-governance-paradigm-in-asia-and-the-pacific.pdf.

management, and societal impacts of AI systems.[13]

At a high level, three overlapping regulatory orientations can be identified within the region: principles-based governance models, hybrid innovation–regulation frameworks, and more centralised or prescriptive regulatory regimes.[14] These categories are not mutually exclusive, nor are they static. Rather, they provide a useful analytical lens for understanding how jurisdictions prioritise flexibility, enforcement, and state oversight in responding to AI-related risks.

2.1 Principles-Based and Guidance-Oriented Models

Several Asia–Pacific jurisdictions have adopted principles-based approaches that emphasise ethical guidance, voluntary compliance, and collaboration with industry. These models typically rely on non-binding instruments, sectoral guidance, and soft-law coordination, rather than comprehensive AI legislation.[15]

Japan remains a leading example of a governance model that places strong emphasis on human-centred principles, coordination, and innovation policy. Its Social Principles of Human-Centric AI, first articulated in 2019, frame AI governance around human dignity, diversity, inclusion, sustainability, and societal well-being.[16] Since 2025, this guidance-oriented architecture has been supplemented by a framework statute governing the promotion and national coordination of AI research, development, and use. Japan therefore remains comparatively light-touch in regulatory technique, but it can no longer be characterised as relying on non-binding principles alone.[17] The continuity lies in its preference for national coordination,

---

[13] Calvino, F., H. Costa and D. Haerle (2026), "Digital technology diffusion in the age of AI: Cross-country evidence from microdata", *OECD Science, Technology and Industry Working Papers*, No. 2026/01, OECD Publishing, Paris, https://doi.org/10.1787/ebc2debe-en.

[14] OECD (2025), "AI and the future of social protection in OECD countries", *OECD Artificial Intelligence Papers*, No. 42, OECD Publishing, Paris, https://doi.org/10.1787/7b245f7e-en.

[15] Lee, J. (2024), *The Landscape of AI Regulation in the Asia-Pacific*, LawTech Asia, https://lawtech.asia/the-landscape-of-ai-regulation-in-the-asia-pacific/

[16] Cabinet Secretariat, Government of Japan, Social Principles of Human-Centric AI (2019) https://www.cas.go.jp/jp/seisaku/jinkouchinou/pdf/humancentricai.pdf.

[17] OECD (2019), *Artificial Intelligence in Society*, OECD Publishing, Paris,

standards, guidance, and international engagement rather than an EU-style cross-sector risk classification regime.

Singapore has pursued a similarly pragmatic and highly operationalised model. Its Model AI Governance Framework, first developed for conventional AI and subsequently extended to generative AI, provides detailed guidance on accountability, transparency, explainability, human oversight, and ecosystem-level risk management.[18] The framework is complemented by practical testing and assurance tools such as AI Verify.[19] In 2026, Singapore further extended this approach through a Model AI Governance Framework for Agentic AI, illustrating its preference for updating operational governance instruments as new technical capabilities emerge.

New Zealand also falls broadly within this category, with a distinctive public-sector emphasis. The Algorithm Charter for Aotearoa New Zealand focuses on transparency, accountability, risk assessment, human oversight, and public confidence in significant algorithmic decision-making by government agencies.[20] The Charter also includes a partnership commitment to embed a Te Ao Māori perspective consistent with Treaty principles, while expressly recognising that Māori Data Sovereignty raises broader issues requiring separate consideration. This distinction is important: the Charter engages Indigenous perspectives, but it does not itself purport to resolve Māori data sovereignty.

https://doi.org/10.1787/eedfee77-en.

18 Infocomm Media Development Authority (IMDA), Model AI Governance Framework for Generative AI (2024), https://www.imda.gov.sg/resources/press-releases-factsheets-and-speeches/press-releases/2024/public-consult-model-ai-governance-framework-genai; IMDA, Model AI Governance Framework for Agentic AI (20 May 2026), https://www.imda.gov.sg/-/media/imda/files/about/emerging-tech-and-research/artificial-intelligence/mgf-for-agentic-ai.pdf.

19 AI Verify Foundation, "What is AI Verify?", https://aiverifyfoundation.sg/what-is-ai-verify/.

20 New Zealand Government, Algorithm Charter for Aotearoa New Zealand (2020, current online guidance), https://data.govt.nz/toolkit/data-ethics/government-algorithm-transparency-and-accountability/algorithm-charter/.

2.2 Hybrid Innovation–Regulation Frameworks

Other jurisdictions in the region have adopted hybrid models that combine principles-based guidance with emerging legal infrastructure, regulatory sandboxes, or sector-specific obligations. These approaches seek to balance innovation promotion with more formal mechanisms for risk mitigation and accountability.[21] South Korea exemplifies this hybrid orientation. Its governance framework now combines an innovation-led national strategy with a binding statutory layer under the Framework Act on the Development of Artificial Intelligence and Establishment of Trustworthiness, which entered into force on 22 January 2026.[22] The Act establishes national governance arrangements and introduces differentiated obligations, including transparency requirements and provisions directed to high-impact and generative AI.[23] South Korea therefore illustrates a hybrid model in a stronger sense than before: innovation promotion remains central, but it is accompanied by AI-specific statutory duties and implementing guidance.

India represents a different variant of the hybrid model. Its National Strategy for Artificial Intelligence, released by NITI Aayog, prioritises economic development, public service delivery, and inclusive growth under the banner of "AI for All". Rather than adopting a comprehensive AI statute, India has developed broader digital governance infrastructure, including data protection and digital public infrastructure, that shapes the environment in which AI systems are developed and deployed. Its

---

21 OECD (2024), "Regulatory approaches to Artificial Intelligence in finance", *OECD Artificial Intelligence Papers*, No. 24, OECD Publishing, Paris, https://doi.org/10.1787/f1498c02-en.

22 Ministry of Science and ICT, Republic of Korea, "A New Chapter in the Age of AI: Basic Act on AI Passed at National Assembly Plenary Session" (26 December 2024), https://www.msit.go.kr/eng/bbs/view.do?bbsSeqNo=42&mId=4&mPid=2&nttSeqNo=1071&sCode=eng.

23 Ministry of Science and ICT, Republic of Korea, "The AI Basic Act Comes into Force to Lay the Foundation for Korea to Become a Global AI Powerhouse" (22 January 2026), https://www.msit.go.kr/eng/bbs/view.do?bbsSeqNo=42&mId=4&mPid=2&nttSeqNo=1214&sCode=eng; Ministry of Science and ICT, "MSIT Releases Guidelines on Ensuring AI Transparency" (2026), https://www.msit.go.kr/eng/bbs/view.do?bbsSeqNo=42&mId=4&nttSeqNo=1215&sCode=eng.

approach is therefore better understood in terms of developmental priorities and regulatory sequencing than as an absence of governance ambition.[24][25]

2.3 Centralised and Prescriptive Regulatory Approaches

At the more prescriptive end of the spectrum lies China, which has developed a centralised and application-specific model of AI governance. In contrast to jurisdictions that rely predominantly on voluntary principles or general law, China has introduced binding measures for particular classes of algorithmic and generative AI services, alongside broader rules on data, cybersecurity, and online content.[26]

The Provisions on the Administration of Algorithmic Recommendation Services impose obligations on providers of recommendation algorithms relating to transparency, user rights, and content governance.[27] More recently, the Interim Measures for the Management of Generative Artificial Intelligence Services extend regulatory oversight to generative AI systems, incorporating requirements related to data sourcing, content moderation, and security assessment.[28] These instruments operate within a broader governance architecture that integrates AI regulation with national security, social stability, and information control. China's model is often contrasted with liberal democratic approaches, but it nonetheless exerts significant regional

---

[24] NITI Aayog, National Strategy for Artificial Intelligence: #AIforAll (2018) https://www.niti.gov.in/sites/default/files/2023-03/National-Strategy-for-Artificial-Intelligence.pdf.

[25] Ministry of Electronics and Information Technology (India), Digital Personal Data Protection Act 2023 https://www.meity.gov.in/static/uploads/2024/06/2bf1f0e9f04e6fb4f8fef35e82c42aa5.pdf.

[26] Creemers, Rogier, China's Social Credit System: An Evolving Practice of Control (2018), https://doi.org/10.2139/ssrn.3175792

[27] Office of the Central Cyberspace Affair Commission, Cyberspace Administration of China, Provisions on the Administration of Algorithmic Recommendation Services 互联网信息服务算法推荐管理规定 (2021) https://www.cac.gov.cn/2022-01/04/c_1642894606364259.htm.

[28] Office of the Central Cyberspace Affair Commission, Cyberspace Administration of China, Interim Measures for the Management of Generative Artificial Intelligence Services **生成式人工智能服务管理暂行办法** (2023) https://www.cac.gov.cn/2023-07/13/c_1690898327029107.htm.

influence. Its regulatory choices affect cross-border AI supply chains, platform governance, and global standard-setting debates, making it an unavoidable comparator in any regional analysis.

2.4 Regional Challenges and Fragmentation

As a result, regional convergence is more likely to occur through functional alignment, such as shared risk taxonomies, interoperability standards, and voluntary coordination, rather than through binding regional regulation.[29] This context provides the backdrop against which Australia's AI governance model must be understood. This fragmentation is not merely a function of regulatory immaturity, but reflects deeper asymmetries in institutional capacity, enforcement capability, and political priorities across jurisdictions. While some economies possess well-resourced regulators and mature legal systems capable of adapting existing frameworks to AI-related risks, others continue to prioritise infrastructure development and innovation uptake over formal governance consolidation.

A key challenge lies in the mismatch between the transboundary nature of AI systems and the predominantly domestic orientation of regulatory authority. AI technologies are frequently developed, deployed, and scaled across borders, yet governance responses remain largely confined within national legal frameworks. This creates practical difficulties for accountability, particularly where AI-driven decision-making affects individuals or markets across multiple jurisdictions with differing regulatory expectations. Enforcement capacity further exacerbates these challenges. Even where ethical principles or governance frameworks are well articulated, uneven institutional capability limits their practical impact. Voluntary guidance mechanisms depend heavily on organisational incentives, regulatory credibility, and cultural norms of compliance, which vary significantly across the region. As a result, similarly framed principles may produce markedly different governance

[29] Andre Wirjo, Sylwyn Calizo Jr., Glacer Niño Vasquez, and Emmanuel A. San Andres, *Artificial Intelligence in Economic Policymaking* (2022), APEC Policy Support Unit, Policy Brief No. 52 APEC https://www.apec.org/docs/default-source/publications/2022/11/artificial-intelligence-in-economic-policymaking/222_psu_artificial-intelligence-in-economic-policymaking.pdf

outcomes in practice.

In addition, the Asia–Pacific region is characterised by divergent conceptions of the role of the state, the status of individual rights, and the relationship between innovation policy and public authority. Some jurisdictions place greater emphasis on procedural accountability, reviewability, and regulatory restraint, while others integrate AI governance more directly with industrial policy, security, or information governance. These differences complicate prospects for regulatory harmonisation and constrain the development of shared enforcement mechanisms. Taken together, these structural features suggest that regional AI governance is unlikely to evolve through formal legal convergence in the near term. Instead, alignment is more plausibly achieved through functional coordination, including shared risk taxonomies, interoperable technical standards, and policy dialogue that enables mutual understanding without requiring uniform regulatory outcomes. This context provides the foundation for examining how Australia's governance model operates within, and responds to, this fragmented regional environment.

3. AUSTRALIA'S AI GOVERNANCE ARCHITECTURE

Australia's approach to AI governance is characterised by a deliberate preference for principles-first regulation, supported by the adaptive application of existing legal and regulatory frameworks rather than a single, comprehensive AI statute.[30] This governance architecture reflects broader features of Australia's regulatory tradition, including reliance on independent regulators, sector-specific enforcement, and incremental policy development informed by consultation and evidence-gathering.[31] Australia's AI governance framework was initially articulated through the release of the AI Ethics Principles in 2019.[32] The Principles, comprising values such as human-centredness, fairness, transparency, accountability, privacy protection, and

[30] Australian Government, Department of Industry, Science and Resources, Australia's AI Ethics Principles (2019) https://www.industry.gov.au/publications/australias-ai-ethics-principles.

[31] Productivity Commission, Australian Government, Harnessing Data and Digital Technology, Inquiry Report (2025), https://www.pc.gov.au/inquiries/completed/data-digital.

[32] Australian Government, Department of Industry, Science and Resources, Australia's AI Ethics Principles (2019) https://www.industry.gov.au/publications/australias-ai-ethics-principles.

contestability, were intended to guide both public and private sector development and deployment of AI systems. While non-binding, they established a normative baseline against which AI practices could be assessed across sectors.

The reliance on ethical principles rather than statutory obligations has attracted both praise and criticism. Proponents argue that principles-based governance preserves regulatory flexibility, encourages innovation, and allows governance norms to evolve alongside rapidly changing technologies.[33] Critics, however, have raised concerns about enforceability, regulatory certainty, and the risk of uneven compliance in the absence of legal sanctions.[34] These debates have shaped subsequent policy developments in Australia's AI governance trajectory. Building on the AI Ethics Principles, the Australian Government undertook consultations from 2023 on safe and responsible AI, including proposals for mandatory guardrails in high-risk settings.[35] By late 2025, the policy architecture had shifted toward implementation through the National AI Plan and the Guidance for AI Adoption, which translated earlier safety principles into more operational governance practices.

The resulting policy trajectory has not produced a comprehensive Australian AI Act. Instead, the National AI Plan, released in December 2025, combines industry and capability objectives with a commitment to keeping Australians safe, while the Guidance for AI Adoption provides organisations with practical governance expectations.[36] This incremental approach preserves the central role of existing law and

---

[33] OECD, Regulatory Approaches to Artificial Intelligence in Finance, OECD Artificial Intelligence Papers No 24 (2024), https://doi.org/10.1787/f1498c02-en.

[34] Wachter, S, Mittelstadt, B and Russell, C, 'Why fairness cannot be automated: Bridging the gap between EU non-discrimination law and AI' (2021) 41 Computer Law & Security Review, https://www.sciencedirect.com/science/article/pii/S0267364921000406.

[35] Department of Industry, Science and Resources, Australian Government, Supporting Responsible AI: Discussion Paper (2023), https://consult.industry.gov.au/supporting-responsible-ai; Department of Industry, Science and Resources, Introducing Mandatory Guardrails for AI in High-Risk Settings: Proposals Paper (2024), https://consult.industry.gov.au/ai-mandatory-guardrails.

[36] Australian Government, Department of Industry, Science and Resources, National AI Plan (2 December 2025), https://www.industry.gov.au/publications/national-ai-plan; National Artificial Intelligence Centre, Guidance for AI Adoption (21 October 2025), https://www.industry.gov.au/publications/guidance-for-ai-adoption.

regulators while making AI governance more explicit at the level of organisational practice.[37]The establishment of an Australian AI Safety Institute further marks a move toward dedicated institutional capability without displacing the sectoral legal framework.

Table 1. Australia's de facto AI regulation: how soft-law frameworks and sectoral legal hooks translate AI risks into auditable accountability mechanisms.

Instrument / regulator	**Legal hook**	**AI-relevant risk**	**Assurance and accountability mechanism**	**Illustrative application**
Privacy Act 1988 (Cth) / OAIC	APPs (collection, use, disclosure, security, access/correction); from 10 Dec 2026, APP 1 automated-decision transparency obligations; complaint and investigation powers	Training and inference involving personal information; automated profiling; data quality and security	PIAs, privacy notices, data governance, access/correction processes, ADM transparency disclosures, breach response and regulator enforcement	Use of AI to triage service delivery must manage data minimisation, purpose limitation, and security controls
Australian Consumer Law and competition law / ACCC	Misleading or deceptive conduct and substantiation expectations; unfair practices; market power and competition scrutiny	Vendor claims about accuracy, bias, explainability; consumer-facing automated decisions; dark patterns	Procurement substantiation, audit rights, performance representations and remedies, ongoing monitoring	Contract terms requiring evidence for accuracy and bias claims, and enabling independent evaluation
Online Safety Act / eSafety Commissioner	Safety duties, takedown and regulatory powers, platform-	Content moderation, recommender systems, generative	Transparency, incident response, platform accountability processes,	AI-assisted moderation should retain review trails and escalation pathways for

[37] Australian Government, Department of Industry, Science and Resources, "Australia to establish new institute to strengthen AI safety" (25 November 2025), https://www.industry.gov.au/news/australia-establish-new-institute-strengthen-ai-safety.

Instrument / regulator	Legal hook	AI-relevant risk	Assurance and accountability mechanism	Illustrative application
	focused compliance frameworks	misinformation and harassment risks	reporting and enforcement	high-impact decisions
Anti-discrimination frameworks / AHRC and courts	Direct and indirect discrimination standards; reasonableness and proportionality tests	Disparate impacts in hiring, credit, housing, and service access decisions	Bias testing, justification records, adjustments and review processes, complaint handling	Documented evaluation and justification for model features and thresholds affecting protected groups
Administrative law (ADJR-style reviewability) / tribunals and courts	Lawfulness, procedural fairness, rationality; reasons and record-making expectations	Rights-affecting decisions relying on AI outputs; opaque automation leading to thin records	Reasons-ready files, minimum reviewable trace, contestability and review pathways	Case file captures input/output, human checks, overrides, and the intelligible justification for reliance
Public-sector AI governance / DTA, accountable authorities and ANAO	Policy for the Responsible Use of AI in Government; accountable authority duties; internal controls; performance and risk management expectations	Whole-of-agency and use-case risk; procurement governance; assurance over internally developed and outsourced AI systems	AI use-case registers, accountable owners, impact assessments, transparency statements, risk-based assurance, internal audit and lifecycle monitoring	Commonwealth agencies assess in-scope AI use cases before deployment, maintain internal registers, allocate accountability, and publish transparency statements

3.1 Sectoral Regulation as De Facto AI Law

Although Australia lacks a dedicated AI statute, a significant portion of AI governance is effected through the application of existing legal frameworks. These operate as de facto AI regulation, shaping the design, deployment, and accountability of

AI systems across sectors.[38]

Privacy law plays a central role. The Privacy Act 1988 (Cth), administered by the Office of the Australian Information Commissioner (OAIC), governs the collection, use, disclosure, security, access, and correction of personal information and therefore applies directly to many AI-driven data practices.[39] The OAIC has issued guidance on privacy and commercially available AI, while amendments enacted through the Privacy and Other Legislation Amendment Act 2024 introduce a specific transparency obligation for certain automated decision-making.[40] From 10 December 2026, APP entities using personal information in automated decisions that may significantly affect individuals' rights or interests will be required to disclose prescribed information about those practices in their privacy policies. This development illustrates how general privacy law is being adapted to AI-related decision processes without creating a standalone AI statute.

Consumer protection and competition law, enforced by the Australian Competition and Consumer Commission (ACCC), also intersect with AI governance. The ACCC's work on digital platforms and algorithmic practices highlights risks associated with misleading conduct, unfair practices, and market concentration arising from AI-driven systems.[41] These regulatory interventions underscore the capacity of general competition law to address AI-related harms without bespoke legislation. Online safety regulation, overseen by the eSafety Commissioner under the Online

---

[38] Department of Industry, Science and Resources, Australian Government, National AI Plan: Keep Australians Safe, https://www.industry.gov.au/publications/national-ai-plan/keep-australians-safe

[39] Office of the Australian Information Commissioner, Guidance on Privacy and the Use of Commercially Available AI Products (2025), https://www.oaic.gov.au/privacy/privacy-guidance-for-organisations-and-government-agencies/guidance-on-privacy-and-the-use-of-commercially-available-ai-products.

[40] Office of the Australian Information Commissioner, Consultation on Guidance for Transparency in Automated Decision-Making (18 May 2026), https://www.oaic.gov.au/engage-with-us/consultations/consultation-on-guidance-for-transparency-in-automated-decision-making.

[41] Australian Competition and Consumer Commission, *Digital Platform Services Inquiry – Final Report* (2025) https://www.accc.gov.au/about-us/publications/serial-publications/digital-platform-services-inquiry-2020-25-reports/digital-platform-services-inquiry-final-report-march-2025.

Safety Act 2021 (Cth), has increasing relevance for AI systems involved in content moderation, recommendation algorithms, and generative content.[42] While not designed specifically for AI, the Act imposes obligations that directly affect AI-enabled platforms, particularly where algorithmic systems amplify harmful content.

In addition, anti-discrimination law and administrative law principles provide avenues for challenging biased or opaque automated decision-making, particularly in public sector contexts.[43] Together, these sectoral regimes create a multi-layered governance environment in which AI is regulated indirectly but substantively.

Australia's AI governance architecture remains institutionally dispersed, with responsibility distributed across regulators, departments, and whole-of-government policy bodies. That dispersion is now partly counterbalanced in the Commonwealth public sector by the Policy for the Responsible Use of AI in Government. The updated policy requires agencies to maintain internal registers of in-scope AI use cases, assign accountable owners, and complete impact assessments before deployment, alongside transparency and accountability requirements.[44] By March 2026, a central register brought together Commonwealth AI transparency statements, making public-sector AI governance more visible and reviewable. These mechanisms are administrative rather than legislative, but they materially strengthen ex ante governance within government.

Recent policy developments respond more directly to these coordination and capability challenges. The Guidance for AI Adoption provides a common govern-

---

[42] Online Safety Act 2021 (Cth); eSafety Commissioner, Online Safety Legislation https://www.aph.gov.au/Parliamentary_Business/Bills_LEGislation/Bills_Search_Results/Result?bId=r6680.

[43] Australian Human Rights Commission, Human Rights and Technology Final Report (2021) https://humanrights.gov.au/our-work/rights-and-freedoms/human-rights-and-technology-final-report.

[44] Digital Transformation Agency, "AI Policy Update: Strengthening Responsible Use across Government" (12 January 2026), https://www.dta.gov.au/articles/ai-policy-update-strengthening-responsible-use-across-government; Digital Transformation Agency, "New Central Register of AI Transparency Statements for Commonwealth Entities" (31 March 2026), https://www.dta.gov.au/articles/new-central-register-ai-transparency-statements-commonwealth-entities.

ance vocabulary for organisations, while Australia's AI Safety Institute now supplies dedicated technical capability to analyse and test advanced AI risks and support regulators and agencies. Together with the Commonwealth public-sector policy, these developments make Australia's architecture more institutionalised than a simple "soft-law" description suggests, although coordination across legal regimes remains an enduring challenge.[45]

3.2 Strengths and Limitations of Australia's Approach

Australia's AI governance model offers several strengths. It preserves regulatory agility, leverages established institutional expertise, and allows legal duties to attach to the function and impact of an AI system rather than to the technology label alone. The more recent addition of national guidance and dedicated safety capability also addresses some of the weaknesses of a purely principles-based model. Australia's alignment with OECD principles and international standards further facilitates engagement with diverse regional partners. The model nonetheless retains significant limitations. Australia still lacks a generally applicable set of binding AI-specific duties, and the interaction of multiple sectoral regimes can create uncertainty, uneven coverage, and coordination costs. The central policy question is therefore no longer simply whether Australia should move from soft law to hard law, but how far common AI-specific expectations should supplement existing legal duties without duplicating them.[46]

In practice, compliance obligations are triggered not by the use of AI as such, but by the *function* the system performs within a regulated domain. For example, AI-driven credit assessment tools are subject to consumer protection and anti-discrimi-

---

[45] National Artificial Intelligence Centre, Guidance for AI Adoption (21 October 2025), https://www.industry.gov.au/publications/guidance-for-ai-adoption; Australian Government, Department of Industry, Science and Resources, Australia's AI Safety Institute (2026), https://www.industry.gov.au/science-technology-and-innovation/technology/artificial-intelligence/ai-safety-institute.

[46] Australian Law Reform Commission, Corporate Criminal Responsibility (ALRC Report 136, 2020) https://www.alrc.gov.au/wp-content/uploads/2020/05/ALRC-CCR-Final-Report-websml.pdf.

nation norms, while biometric identification technologies engage privacy and surveillance safeguards. This functional orientation reflects a regulatory logic that focuses on outcomes rather than technological form, allowing existing legal principles to be applied to novel AI applications. However, this approach places a significant interpretive burden on both regulators and regulated entities. Organisations must navigate overlapping obligations across privacy, consumer, competition, and online safety regimes, often without clear ex ante guidance on how AI-specific risks should be assessed. Regulatory clarity therefore depends heavily on guidance, enforcement signalling, and institutional coordination, rather than on statutory prescription. This reinforces the importance of regulatory capability and consistency in sustaining Australia's governance model.

4. COMPARATIVE PERSPECTIVES: AUSTRALIA AND KEY ASIA–PACIFIC JURISDICTIONS

The comparative analysis that follows is not intended to rank jurisdictions or identify a preferred regulatory model. Rather, it uses comparison to illuminate how governance choices reflect institutional capacity, political values, regulatory sequencing, and risk tolerance. Each comparator highlights a different dimension of Australia's approach, including normative alignment, administrative pragmatism, statutory institutionalisation, developmental sequencing, and centralised control, thereby clarifying Australia's distinctive position within the regional governance spectrum. This comparative lens allows Australia's governance model to be understood relationally rather than in isolation, revealing both its strengths and its constraints when situated alongside alternative regulatory trajectories in the Asia–Pacific.

This section situates Australia's AI governance model within a comparative Asia–Pacific context by examining selected approaches adopted in Japan, Singapore, South Korea, New Zealand, India, and China. The comparison is structured around four dimensions: regulatory philosophy, institutional architecture, the relative reliance on hard law versus soft-law instruments, and approaches to enforcement and accountability. Given the breadth of jurisdictions considered, the objective of this comparative exercise is not to provide jurisdiction-specific doctrinal analysis, but rather to map

the regional governance landscape and to highlight key points of convergence and divergence in regulatory design choices across the region. The focus is therefore on identifying overarching patterns, strategic orientations, and governance trade-offs, rather than on granular regulatory detail.

Viewed through this comparative lens, Australia's approach to AI governance becomes more intelligible. While Australia shares a number of normative commitments with other Asia–Pacific jurisdictions, particularly an emphasis on human-centred values, transparency, and accountability, it diverges in important respects, most notably in its institutional arrangements, regulatory sequencing, and relatively high tolerance for legal indeterminacy.

4.1 Australia and Japan: Normative Alignment, Institutional Divergence

Australia and Japan exhibit strong normative alignment in their articulation of AI governance principles. Both jurisdictions emphasise human-centred AI, social trust, and alignment with international norms, particularly those developed by the OECD. [47] However, they diverge in how these principles are institutionalised.

Japan's Social Principles of Human-Centric AI remain embedded within a broader national innovation strategy that integrates government, industry, and academia. Since 2025, however, Japan has also placed AI policy on a statutory footing through the Act on the Promotion of Research and Development and the Utilization of Artificial Intelligence-Related Technology. The Act was fully brought into force in September 2025 and establishes national coordination and planning mechanisms while retaining a promotion-oriented, comparatively light-touch approach. Australia, by contrast, continues to rely more heavily on independent regulators and existing legal

[47] OECD, *OECD Principles on Artificial Intelligence* (2024) https://www.oecd.org/en/topics/sub-issues/ai-principles.html

frameworks to give practical effect to accountability norms.[48-49]

4.2 Australia and Singapore: Pragmatism and Operationalisation

Singapore offers one of the most operationalised AI governance models in the region. Its Model AI Governance Framework and related generative-AI framework translate high-level principles into organisational practices, while AI Verify provides testing and assurance infrastructure.[50] The 2026 Model AI Governance Framework for Agentic AI extends this iterative approach to systems capable of more autonomous action. Singapore's model is therefore characterised not merely by voluntary guidance, but by a sustained effort to convert governance principles into usable technical and organisational tools.

Australia shares Singapore's pragmatic orientation but differs in institutional design. Singapore has developed bespoke AI governance instruments and testing infrastructure, whereas Australia has generally sought to integrate AI governance into existing regulators and legal regimes, supplemented by cross-cutting guidance. The difference is therefore less one of regulatory ambition than of administrative architecture and the degree to which AI-specific governance is centralised.[51]

---

48 Cabinet Secretariat, Government of Japan, Social Principles of Human-Centric AI (2019), https://www.cas.go.jp/jp/seisaku/jinkouchinou/pdf/humancentricai.pdf; Cabinet Office, Government of Japan, Act on the Promotion of Research and Development and the Utilization of Artificial Intelligence-Related Technology (promulgated 4 June 2025; fully in force 1 September 2025), https://www8.cao.go.jp/cstp/ai/ai_act/ai_act.html.

49 Australian Government, Department of Industry, Science and Resources, National AI Plan (2 December 2025), https://www.industry.gov.au/publications/national-ai-plan; National Artificial Intelligence Centre, Guidance for AI Adoption (21 October 2025), https://www.industry.gov.au/publications/guidance-for-ai-adoption.

50 AI Verify Foundation, "What is AI Verify?", https://aiverifyfoundation.sg/what-is-ai-verify; Infocomm Media Development Authority, Model AI Governance Framework for Agentic AI (20 May 2026), https://www.imda.gov.sg/-/media/imda/files/about/emerging-tech-and-research/artificial-intelligence/mgf-for-agentic-ai.pdf.

51 Productivity Commission, *Harnessing data and digital technology* (2025) https://assets.pc.gov.au/2025-12/data-digital_0.pdf.

4.3 Australia and South Korea: Regulatory Readiness and State Capacity

South Korea now provides the clearest statutory contrast with Australia among the jurisdictions considered here. The AI Basic Act, in force from January 2026, combines innovation promotion with national governance structures and AI-specific obligations, including transparency requirements and differentiated treatment of high-impact and generative AI. This gives South Korea a more formalised ex ante governance architecture while retaining a policy emphasis on technological development.[52]

Australia remains more institutionally dispersed by comparison. Although it has added national guidance and dedicated AI safety capability, legal accountability continues to arise primarily through general and sectoral regimes rather than a single AI-specific framework statute. The contrast therefore concerns where legal obligations are located and coordinated: South Korea has consolidated important AI-specific expectations in legislation, while Australia continues to distribute them across existing legal domains and administrative mechanisms.

4.4 Australia and New Zealand: Shared Values, Different Regulatory Emphases

Australia and New Zealand share common legal traditions and commitments to transparency and accountability. New Zealand's Algorithm Charter, however, gives unusually explicit attention to public-sector algorithmic governance, including risk assessment, explainability, human oversight, and a partnership commitment to Te Ao Māori perspectives consistent with Treaty principles.[53]The Charter simultane-

[52] Ministry of Science and ICT, Republic of Korea, "The AI Basic Act Comes into Force to Lay the Foundation for Korea to Become a Global AI Powerhouse" (22 January 2026), https://www.msit.go.kr/eng/bbs/view.do?bbsSeqNo=42&mId=4&mPid=2&nttSeqNo=1214&sCode=eng.

[53] New Zealand Government, Algorithm Charter for Aotearoa New Zealand (2020, current online guidance), https://data.govt.nz/toolkit/data-ethics/government-algorithm-transparency-and-

ously acknowledges that Māori Data Sovereignty requires separate consideration, illustrating both the reach and the limits of a general algorithm-governance instrument.

Australia's AI governance discourse has placed greater emphasis on economic opportunity, safety, and risk management than on explicit Indigenous governance frameworks. New Zealand's Algorithm Charter therefore illustrates a different way in which local constitutional and cultural commitments can shape public-sector algorithmic governance even without a comprehensive AI statute. The comparison underscores that principles-based governance is not normatively neutral; its content depends on the institutional and social context in which principles are operationalised.

4.5 Australia and India: Developmental Priorities and Regulatory Sequencing

India's AI governance approach is shaped by developmental imperatives, the scale of digital transformation, and a policy preference for sequencing infrastructure, inclusion, and governance reforms. The AI for All strategy prioritises public service delivery, economic inclusion, and technological capability, with AI-specific governance developing alongside broader data and digital regulation.[54] Australia, by contrast, operates within a mature regulatory environment with higher baseline expectations of accountability and rights protection. While India has moved rapidly to develop digital public infrastructure, including data governance frameworks, it has deferred comprehensive AI regulation.[55] Australia's challenge is different: managing AI risks within an already dense regulatory ecosystem without overburdening innovation.

---

accountability/algorithm-charter/.

[54] NITI Aayog, National Strategy for Artificial Intelligence: #AIforAll (2018) https://www.niti.gov.in/sites/default/files/2023-03/National-Strategy-for-Artificial-Intelligence.pdf.

[55] Ministry of Electronics and Information Technology (India), *The Digital Personal Data Protection Act* (2023) https://www.meity.gov.in/static/uploads/2024/06/2bf1f0e9f04e6fb4f8fef35e82c42aa5.pdf

4.6 Australia and China: Contrasting Governance Logics

The contrast between Australia and China is the most pronounced. China's AI governance model is characterised by centralised authority, binding obligations, and integration with broader objectives relating to social governance, security, and information control.[56] Regulatory instruments targeting algorithmic recommendation services and generative AI impose explicit compliance requirements, including content controls, data governance obligations, and security assessments.[57]

Australia's legal-institutional logic differs substantially. Accountability is distributed across independent regulators, general and sectoral law, administrative review, and rights-protective frameworks rather than integrated into a centralised system of technology and information control. Nonetheless, China's regulatory choices affect Australia indirectly through cross-border platforms, supply chains, data practices, and international standards debates. China therefore remains a necessary comparator even where deeper normative convergence is limited.[58]

4.7 Comparative Insights

Several comparative insights emerge from this analysis. First, the absence of a single optimal model of AI governance across the Asia–Pacific reflects the extent to which regulatory approaches are shaped by domestic institutional capacity, political culture, legal tradition, and strategic priorities. Rather than converging on a uniform regulatory template, jurisdictions calibrate their governance frameworks in light of

---

[56] Creemers, Rogier, China's Social Credit System: An Evolving Practice of Control (2018), https://doi.org/10.2139/ssrn.3175792

[57] Office of the Central Cyberspace Affair Commission, Cyberspace Administration of China, Interim Measures for the Management of Generative Artificial Intelligence Services **生成式人工智能服务管理暂行办法** (2023) https://www.cac.gov.cn/2023-07/13/c_1690898327029107.htm.

[58] Australian Human Rights Commission, *Human Rights and Technology Final Report* (2021) https://humanrights.gov.au/resource-hub/by-resource-type/publications/final-report-human-rights-and-technology

differing risk tolerances, innovation objectives, and assumptions about the appropriate role of the state in technology regulation. This diversity underscores the limitations of transplanting governance models across contexts without regard to underlying institutional and normative conditions.

Second, Australia occupies an intermediate but increasingly institutionalised position within this regional spectrum. It combines non-binding principles and implementation guidance with enforceable duties drawn from existing law, while adding dedicated national policy and safety capability. This configuration preserves flexibility but also exposes tensions around coherence, ex ante certainty, and the allocation of responsibility where AI systems cross sectoral boundaries.

Third, the comparative analysis suggests that regional convergence in AI governance is more likely to emerge through shared principles, technical standards, and functional forms of coordination than through formal harmonisation of legislation. Given the heterogeneity of legal systems and governance traditions in the region, soft alignment mechanisms, such as interoperable frameworks, mutual recognition practices, and collaborative regulatory experimentation, offer a more realistic pathway for managing cross-border AI risks and facilitating cooperation. Taken together, these insights provide the conceptual foundation for the next section, which examines Australia's potential role as a bridging jurisdiction in regional AI governance, one capable of translating between divergent regulatory models while contributing to principled, yet flexible, forms of regional coordination.

5. REGIONAL COOPERATION, SOFT CONVERGENCE, AND AUSTRALIA AS A BRIDGING JURISDICTION

The absence of a binding regional framework for AI governance in the Asia–Pacific has not resulted in regulatory vacuum. Instead, the region has witnessed the emergence of soft convergence through multilateral dialogue, technical standards, and functional coordination mechanisms that allow jurisdictions with divergent legal systems and political values to engage without formal harmonisation. This form of convergence privileges interoperability over uniformity and reflects the region's institutional diversity and geopolitical sensitivities.

5.1 Soft Convergence through Multilateral Dialogue

Multilateral forums play a central role in shaping shared governance vocabularies for AI in the Asia–Pacific. APEC has become more explicit in this field. In 2025, APEC leaders endorsed the APEC Artificial Intelligence Initiative for 2026–2030, centred on AI-enabled economic growth, broader participation in AI transformation, capacity building, and investment in resilient AI infrastructure. The Initiative does not harmonise domestic law, but it provides a region-wide platform through which economies can develop shared priorities and practical cooperation.[59] Australia participates in these APEC processes as part of its broader regional digital and technology engagement. APEC's Digital Economy Steering Group and related standards and capacity-building work provide channels through which Australia can contribute to interoperability and policy learning without requiring member economies to adopt uniform legal rules.[60]

Parallel, though more limited, contributions have emerged through the Quadrilateral Security Dialogue (Quad). The Quad's principles on technology design and governance emphasise values-based cooperation among like-minded states, reinforcing expectations around transparency, resilience, and responsible innovation.[61] While not region-wide, these initiatives contribute to norm diffusion and reinforce a shared governance language that influences broader regional discourse.

While soft convergence offers a pragmatic pathway for regional alignment, it is inherently fragile. Its effectiveness depends on sustained political commitment,

---

[59] APEC, APEC Artificial Intelligence (AI) Initiative (2026–2030), endorsed in the 2025 APEC Leaders' Gyeongju Declaration (1 November 2025), https://www.apec.org/meeting-papers/leaders-declarations/2025/2025-apec-leaders--gyeongju-declaration/apec-artificial-intelligence-%28ai%29-initiative-%282026-2030%29.

[60] APEC, Digital Economy Steering Group, https://www.apec.org/groups/committee-on-trade-and-investment/digital-economy-steering-group; APEC, "APEC Economies Advance AI Standards for Trust and Innovation" (3 August 2025), https://www.apec.org/press/news-releases/2025/apec-economies-advance-ai-standards-for-trust-and-innovation.

[61] Quad Leaders' Joint Statement, *Quad Principles on Technology Design, Development, Governance, and Use* (2023) https://www.pm.gov.au/media/quad-leaders-joint-statement

regulatory capability, and mutual trust among participating jurisdictions. In periods of geopolitical tension or rapid technological disruption, these conditions may erode, weakening the normative influence of non-binding frameworks and standards. For Australia, this fragility underscores the importance of reinforcing domestic regulatory legitimacy. Soft-law and standards-based engagement is most credible when supported by demonstrable regulatory capability at home. If domestic expectations remain too diffuse or difficult to enforce, Australia's ability to function as an effective governance intermediary may be weakened by perceptions of regulatory insufficiency.

5.2 Standards and Functional Interoperability

Technical standards constitute a second pillar of soft convergence. International standard-setting bodies, particularly ISO/IEC JTC 1/SC 42, have developed standards addressing AI risk management, data governance, system robustness, and transparency.[62] These standards offer a practical mechanism for coordination across jurisdictions without requiring legislative harmonisation. Australia's engagement with international standards bodies reflects a strategic preference for functional interoperability. Standards provide regulators with reference points for assessing compliance under existing legal frameworks and offer industry a predictable baseline for cross-border operation. In a region characterised by regulatory diversity, standards facilitate coordination while avoiding the political and legal challenges associated with binding regional regulation.

5.3 Australia's Bridging Role in Regional AI Governance

Australia's participation in multilateral forums and standards development illustrates its potential role as a bridging jurisdiction within the Asia–Pacific AI governance landscape. This role does not involve exporting regulatory models or imposing normative leadership. Rather, it involves translation and mediation, connecting

[62] ISO/IEC JTC 1/SC 42, *Artificial Intelligence Standards Programme,* https://www.iso.org/committee/6794475.html.

principles-based governance with enforceable legal frameworks, and aligning innovation-oriented approaches with accountability expectations. Australia's role as a bridging jurisdiction is best understood as a process of governance translation rather than regulatory leadership. Translation involves rendering abstract principles, such as fairness, accountability, and transparency, operational across divergent institutional contexts, without insisting on uniform legal expression. In this sense, Australia's contribution lies in facilitating mutual intelligibility between governance systems, rather than advancing a singular regulatory template.

This translational role is particularly evident in standards-based cooperation and multilateral policy dialogue, where Australia is able to engage with both innovation-oriented and risk-averse jurisdictions. By anchoring engagement in shared principles and functional expectations, Australia can support interoperability while respecting regulatory pluralism. This form of influence is subtle but durable, relying on credibility, institutional trust, and sustained engagement rather than coercive authority.[63]

Australia's domestic AI governance architecture partly supports this positioning. Its combination of principles, sectoral regulation, implementation guidance, and emerging safety capability enables engagement with jurisdictions across the regulatory spectrum. At the same time, the credibility of any bridging role depends on Australia demonstrating that flexibility does not come at the expense of effective accountability.[64] In this sense, Australia functions less as a regulatory leader and more as a governance intermediary, facilitating dialogue, interoperability, and trust across diverse systems. This intermediary role aligns with broader theories of middle-power influence, which emphasise coalition-building, norm translation, and institutional brokerage rather than rule-making dominance.

5.4 Strategic Implications

The bridging role is not without limits. Soft convergence mechanisms lack en-

---

[63] Slaughter, A-M, *A New World Order* (Princeton University Press, 2004) P255

[64] Australian Human Rights Commission, *Human Rights and Technology Final Report* (2021) https://humanrights.gov.au/resource-hub/by-resource-type/publications/final-report-human-rights-and-technology

forceability and rely heavily on voluntary compliance and institutional capacity.[65] Political divergence, strategic competition, and differing conceptions of state authority constrain the scope for deeper regional integration, particularly where AI governance intersects with national security, information control, or social stability.[66] For Australia, reliance on voluntary alignment and standards continues to raise questions about domestic accountability and public trust, notwithstanding the more formal policy architecture introduced since 2025. As AI systems become more pervasive and consequential, pressure may grow for clearer and, in selected contexts, binding AI-specific obligations to complement sectoral law. The challenge is to improve certainty and protection without creating duplicative regulation or undermining the interoperability advantages of Australia's current approach.

The Asia–Pacific AI governance landscape is likely to continue evolving through incremental alignment rather than harmonisation. Within this context, Australia's strategic value lies in its capacity to operate across governance divides, supporting interoperability and normative coherence without insisting on uniformity. This bridging role positions Australia as a credible interlocutor in regional AI governance debates, while also highlighting the importance of continued domestic regulatory development to sustain legitimacy and trust. The balance between soft convergence and legal accountability will remain a defining challenge for Australia, and for middle-power jurisdictions more broadly, in the next phase of AI governance.

6. CONCLUSION

AI governance in the Asia–Pacific region is marked by diversity rather than convergence. Jurisdictions across the region have adopted distinct regulatory approaches shaped by political systems, institutional capacity, economic priorities, and normative commitments. This chapter has shown that, in the absence of binding regional regulation, AI governance in the Asia–Pacific is evolving through soft con-

---

[65] OECD Artificial Intelligence Papers, *AI Policy Observatory: AI Openness A Primer for Policymakers,* (2025) https://www.oecd.org/content/dam/oecd/en/publications/reports/2025/08/ai-openness_958d292b/02f73362-en.pdf

[66] Creemers, Rogier, China's Social Credit System: An Evolving Practice of Control (2018), https://doi.org/10.2139/ssrn.3175792

vergence, via shared principles, technical standards, policy dialogue, and functional interoperability rather than harmonised legal frameworks. Within this landscape, Australia occupies a distinctive position. Its AI governance model remains principles-first and sectoral, but it is no longer accurately described as relying on soft law alone. The National AI Plan, implementation guidance, regulator activity, and dedicated safety capability have progressively formalised the institutional environment around AI while leaving existing legal regimes to carry most enforceable obligations. This combination reflects Australia's continuing effort to balance innovation, accountability, and regulatory burden.

The comparative analysis demonstrates that Australia neither mirrors highly centralised or prescriptive regimes nor relies solely on voluntary self-regulation. It combines guidance with enforceable legal mechanisms in areas such as privacy, consumer protection, competition, online safety, anti-discrimination, and public law. Compared with Japan and Singapore, Australia relies more heavily on distributed legal accountability; compared with South Korea, it has not consolidated AI-specific duties in a framework statute; and compared with China, it operates through a fundamentally different institutional and legal logic.

Importantly, Australia's role in the region is best understood not as that of a regulatory leader or model exporter, but as a bridging jurisdiction. Australia contributes to regional AI governance by translating shared principles into practical governance tools, supporting standards-based interoperability, and facilitating dialogue across regulatory and political divides. Its engagement in multilateral forums such as APEC, participation in international standards development, and commitment to capacity-building illustrate how middle-power jurisdictions can exercise influence through brokerage rather than dominance.

At the same time, the analysis highlights the limits of soft convergence. Voluntary alignment and standards-based coordination cannot substitute for credible domestic accountability where AI systems produce legally or socially consequential harms. For Australia, sustaining credibility as a regional interlocutor will depend on clearer coordination across regulators, technically capable assurance institutions, and visible pathways for review and remedy. For Australia to perform a credible bridging function, three conditions matter. First, domestic coordination must be strong enough to translate principles into consistent regulator expectations and auditable practices

across sectors. Second, Australia needs sustained capability in evaluation, procurement, and assurance so that standards and guidance are backed by measurable compliance and contestability. Third, the legitimacy of soft convergence depends on visible accountability when harms occur, including clear review pathways, transparent explanations, and enforceable remedies.

More broadly, this chapter contributes to understanding how AI governance is likely to develop in regions characterised by legal diversity and geopolitical complexity. The Asia–Pacific experience suggests that incremental alignment, norm translation, and functional interoperability may be more realistic, and more durable, than formal harmonisation. In this context, Australia's experience offers insights into the possibilities and constraints of middle-power governance in an era of fragmented technological regulation. As AI technologies continue to evolve, the challenge for Australia and its regional partners will be to maintain trust, legitimacy, and accountability while preserving space for innovation. Australia's bridging role should therefore be understood as conditional rather than inherent: it depends on the country's capacity to translate principles into auditable practice, connect standards with enforceable legal duties, and engage across divergent regional systems without overstating the degree of normative convergence.

REFERENCES

Andre Wirjo, Sylwyn Calizo Jr., Glacer Niño Vasquez, and Emmanuel A. San Andres, Artificial Intelligence in Economic Policymaking (2022), APEC Policy Support Unit, Policy Brief No. 52 APEC https://www.apec.org/docs/default-source/publications/2022/11/artificial-intelligence-in-economic-policymaking/222_psu_artificial-intelligence-in-economic-policymaking.pdf.

Aneja, U., Gupta, A. and John, S. (2025), Exploring the Potential of AI to Usher in a New Governance Paradigm in Asia and the Pacific, Digital Future Labs- United Nations Development Programme, https://www.undp.org/sites/g/files/zskgke326/files/2025-12/exploring-the-potential-of-ai-to-usher-in-a-new-governance-paradigm-in-asia-and-the-pacific.pdf.

APEC, AI Governance: Why Cooperation Matters (2024) https://www.apec.org/press/features/2024/ai-governance--why-cooperation-matters.

APEC, APEC Artificial Intelligence (AI) Initiative (2026–2030), endorsed in the 2025 APEC Leaders' Gyeongju Declaration (1 November 2025) https://www.apec.org/meeting-papers/leaders-declarations/2025/2025-apec-leaders--gyeongju-declaration/apec-artificial-intelligence-%28ai%29-initiative-%282026-2030%29.

APEC, Digital Economy Steering Group https://www.apec.org/groups/committee-on-trade-and-investment/digital-economy-steering-group.

APEC, "APEC Economies Advance AI Standards for Trust and Innovation" (3 August 2025) https://www.apec.org/press/news-releases/2025/apec-economies-advance-ai-standards-for-trust-and-innovation.

Australian Competition and Consumer Commission, Digital Platform Services Inquiry – Final Report (2025) https://www.accc.gov.au/about-us/publications/serial-publications/digital-platform-services-inquiry-2020-25-reports/digital-platform-services-inquiry-final-report-march-2025.

Australian Government Department of the Prime Minister and Cabinet, Quad Leaders' Summit, Critical and Emerging Technology (2023) https://www.pmc.gov.au/resources/quad-leaders-summit-2023/critical-and-emerging-technology.

Australian Government, Department of Industry, Science and Resources, Australia's AI Ethics Principles (2019) https://www.industry.gov.au/publications/australias-ai-ethics-principles.

Australian Government, Department of Industry, Science and Resources, Australia's AI Safety Institute (2026) https://www.industry.gov.au/science-technology-and-innovation/technology/artificial-intelligence/ai-safety-institute.

Australian Government, Department of Industry, Science and Resources, National AI Plan (2 December 2025) https://www.industry.gov.au/publications/national-ai-plan.

Australian Government, Department of Industry, Science and Resources, "Australia to establish new institute to strengthen AI safety" (25 November 2025) https://www.industry.gov.au/news/australia-establish-new-institute-strengthen-ai-safety.

Australian Human Rights Commission, Human Rights and Technology Final Report (2021) https://humanrights.gov.au/resource-hub/by-resource-type/publications/final-report-human-rights-and-technology.

Australian Law Reform Commission, Corporate Criminal Responsibility (ALRC Report 136, 2020) https://www.alrc.gov.au/wp-content/uploads/2020/05/ALRC-CCR-Final-Report-websml.pdf.

Australian National Audit Office, Administration of the frameworks developed by the Digital Transformation Agency (DTA) for the use of emerging technologies in the public sector (2026) https://www.anao.gov.au/work/performance-audit/administration-of-the-frameworks-developed-the-digital-transformation-agency-dta-the-use-of-emerging.

Cabinet Office, Government of Japan, Act on the Promotion of Research and Development and the Utilization of Artificial Intelligence-Related Technology (promulgated 4 June 2025; fully in force 1 September 2025) https://www8.cao.go.jp/cstp/ai/ai_act/ai_act.html.

Cabinet Secretariat, Government of Japan, Social Principles of Human-Centric AI (2019) https://www.cas.go.jp/jp/seisaku/jinkouchinou/pdf/humancentricai.pdf.

Calvino, F., H. Costa and D. Haerle (2026), "Digital technology diffusion in the age of AI: Cross-

country evidence from microdata", OECD Science, Technology and Industry Working Papers, No. 2026/01, OECD Publishing, Paris, https://doi.org/10.1787/ebc2debe-en.

Creemers, Rogier, China's Social Credit System: An Evolving Practice of Control (2018), https://doi.org/10.2139/ssrn.3175792.

Department of Foreign Affairs and Trade, Australian Government, Australia's Digital Economy Strategy (2022) https://www.dfat.gov.au/sites/default/files/digital-trade-strategy.pdf.

Department of Industry, Science and Resources, Australian Government, National AI Plan: Keep Australians Safe, https://www.industry.gov.au/publications/national-ai-plan/keep-australians-safe.

Department of Industry, Science and Resources, Australian Government, Supporting responsible AI: discussion paper (2023) https://consult.industry.gov.au/supporting-responsible-ai/submission/list.

Digital Transformation Agency, "AI Policy Update: Strengthening Responsible Use across Government" (12 January 2026) https://www.dta.gov.au/articles/ai-policy-update-strengthening-responsible-use-across-government.

Digital Transformation Agency, "New Central Register of AI Transparency Statements for Commonwealth Entities" (31 March 2026) https://www.dta.gov.au/articles/new-central-register-ai-transparency-statements-commonwealth-entities.

eSafety Commissioner, Online Safety Legislation https://www.aph.gov.au/Parliamentary_Business/Bills_LEGislation/Bills_Search_Results/Result?bId=r6680.

Infocomm Media Development Authority (IMDA), AI Verify https://aiverifyfoundation.sg/what-is-ai-verify/.

Infocomm Media Development Authority (IMDA), Model AI Governance Framework for Agentic AI (20 May 2026) https://www.imda.gov.sg/-/media/imda/files/about/emerging-tech-and-research/artificial-intelligence/mgf-for-agentic-ai.pdf.

Infocomm Media Development Authority (IMDA), Model AI Governance Framework for Generative AI (2024), https://www.imda.gov.sg/resources/press-releases-factsheets-and-speeches/press-releases/2024/public-consult-model-ai-governance-framework-genai.

ISO/IEC JTC 1/SC 42, Artificial Intelligence Standards Programme, https://www.iso.org/committee/6794475.html.

Lee, J. (2024), The Landscape of AI Regulation in the Asia-Pacific, LawTech Asia, https://lawtech.asia/the-landscape-of-ai-regulation-in-the-asia-pacific/.

Ministry of Electronics and Information Technology (India), Digital Personal Data Protection Act (2023) https://www.meity.gov.in/static/uploads/2024/06/2bf1f0e9f04e6fb4f8fef35e82c42aa5.pdf.

Ministry of Science and ICT, Republic of Korea, "A New Chapter in the Age of AI: Basic Act on AI Passed at National Assembly Plenary Session" (26 December 2024)

https://www.msit.go.kr/eng/bbs/view.do?bbsSeqNo=42&mId=4&mPid=2&nttSeqNo=1071&sCode=eng.

Ministry of Science and ICT, Republic of Korea, "MSIT Releases Guidelines on Ensuring AI Transparency" (2026) https://www.msit.go.kr/eng/bbs/view.do?bbsSeqNo=42&mId=4&nttSeqNo=1215&sCode=eng.

Ministry of Science and ICT, Republic of Korea, "The AI Basic Act Comes into Force to Lay the Foundation for Korea to Become a Global AI Powerhouse" (22 January 2026) https://www.msit.go.kr/eng/bbs/view.do?bbsSeqNo=42&mId=4&mPid=2&nttSeqNo=1214&sCode=eng.

National Artificial Intelligence Centre, Guidance for AI Adoption (21 October 2025) https://www.industry.gov.au/publications/guidance-for-ai-adoption.

New Zealand Government, Algorithm Charter for Aotearoa New Zealand (2020) https://www.data.govt.nz/toolkit/data-ethics/government-algorithm-transparency-and-accountability/algorithm-charter/.

NITI Aayog, National Strategy for Artificial Intelligence: #AIforAll (2018) https://www.niti.gov.in/sites/default/files/2023-03/National-Strategy-for-Artificial-Intelligence.pdf.

OECD (2019), Artificial Intelligence in Society, OECD Publishing, Paris, https://doi.org/10.1787/eedfee77-en.

OECD (2024), "Regulatory approaches to Artificial Intelligence in finance", OECD Artificial Intelligence Papers, No. 24, OECD Publishing, Paris, https://doi.org/10.1787/f1498c02-en.

OECD (2025), "AI and the future of social protection in OECD countries", OECD Artificial Intelligence Papers, No. 42, OECD Publishing, Paris, https://doi.org/10.1787/7b245f7e-en.

OECD (2025), "How can AI help make better regional development policies in Korea", OECD Regional Development Papers, OECD Publishing, Paris, https://www.oecd.org/content/dam/oecd/en/publications/reports/2025/12/how-can-ai-help-make-better-regional-development-policies-in-korea_2f49d378/be01f52d-en.pdf.

OECD Artificial Intelligence Papers, AI Policy Observatory: AI Openness A Primer for Policymakers, (2025) https://www.oecd.org/content/dam/oecd/en/publications/reports/2025/08/ai-openness_958d292b/02f73362-en.pdf.

OECD, Framework for the Classification of AI Systems (2022) https://www.oecd.org/going-digital/ai/classification-of-ai-systems.htm.

OECD, OECD Principles on Artificial Intelligence (2019, updated 2024) https://oecd.ai/en/ai-principles.

OECD, Regulatory Approaches to Artificial Intelligence in Finance, OECD Artificial Intelligence Papers No 24 (2024), https://doi.org/10.1787/f1498c02-en.

Office of the Australian Information Commissioner, Australian Government, Guidance on privacy and the use of commercially available AI products (2025), https://www.oaic.gov.au/privacy/privacy-guidance-for-organisations-and-government-agencies/guidance-on-privacy-and-the-use-of-commercially-available-ai-products.

Office of the Australian Information Commissioner, Consultation on Guidance for Transparency in Automated Decision-Making (18 May 2026) https://www.oaic.gov.au/engage-with-us/consultations/consultation-on-guidance-for-transparency-in-automated-decision-making.

Office of the Central Cyberspace Affair Commission, Cyberspace Administration of China, Interim Measures for the Management of Generative Artificial Intelligence Services 生成式人工智能服务管理暂行办法 (2023) https://www.cac.gov.cn/2023-07/13/c_1690898327029107.htm.

Office of the Central Cyberspace Affair Commission, Cyberspace Administration of China, Provisions on the Administration of Algorithmic Recommendation Services 互联网信息服务算法推荐管理规定 (2021) https://www.cac.gov.cn/2022-01/04/c_1642894606364259.htm.

Online Safety Act 2021 (Cth).

Productivity Commission, Australian Government, Harnessing data and digital technology Inquiry Report (2025) https://assets.pc.gov.au/2025-12/data-digital_0.pdf.

Productivity Commission, Regulation of New and Emerging Technologies (2023) https://www.pc.gov.au/inquiries-and-research/productivity/.

Quad Leaders' Joint Statement, Quad Principles on Technology Design, Development, Governance, and Use (2023) https://www.pm.gov.au/media/quad-leaders-joint-statement.

Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act), OJ L 2024/1689, 12 July 2024, https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng.

Slaughter, A-M, A New World Order (Princeton University Press, 2004) P255.

UNESCO, Recommendation on the Ethics of Artificial Intelligence (2021) https://www.unesco.org/en/artificial-intelligence/recommendation-ethics.

Wachter, S, Mittelstadt, B and Russell, C, 'Why fairness cannot be automated: Bridging the gap between EU non-discrimination law and AI' (2021) 41 Computer Law & Security Review, https://www.sciencedirect.com/science/article/pii/S0267364921000406.

DIVERGENT PATHS IN AI REGULATION: AUSTRALIA AND BRAZIL AT A MOMENT OF GLOBAL RECALIBRATION

Leonardo Andrade Macedo

Master of Laws and PhD candidate in Law (UFMG)

Federal Prosecutor (MPF)

DOI: https://doi.org/10.59224/dti8.ch12

Abstract: This article examines Australia's evolving approach to artificial intelligence regulation in comparison with Brazil's. After surveying regulatory rationales and strategies, it traces Australia's trajectory from voluntary ethical principles to a recent retreat from binding rules, drawing lessons for Brazil's pending AI bill (PL 2338/2023). It also identifies how Brazil's institutional architecture may inform the Australian debate.

Keywords: regulation; artificial intelligence; Australia–Brazil.

Table of contents: 1. Introduction. 2. Regulating AI: rationales and strategies. 3. Australia's AI regulatory trajectory. 3.1 Ethical principles as soft-law foundations. 3.2 Consultation, risk, and the case for mandatory safeguards. 3.3 Voluntary guardrails: from VAISS to AI6. 3.4 Mandatory guardrails and the technology-neutral pivot. 4. Australia and Brazil in comparative perspective: divergent paths, mutual lessons. 5. Concluding remarks.

1. INTRODUCTION

Although technological innovation has long figured on the regulatory agenda of states, particularly those with deep traditions in science, research and development,

efforts to regulate artificial intelligence (AI) have intensified globally over the past decade. They accelerated further with recent advances in generative AI and the popularisation of these tools among ordinary users.

The "race to AI", premised on the idea that nations leading in this technology will gain decisive advantages in the global market, not only economically and socially, but, above all, militarily and geopolitically, has been accompanied by a parallel "race to AI regulation". This new race is based on the view that an adequate regulatory framework can offer legal certainty to innovators and strengthen public trust in the technology, thereby producing a competitive advantage of its own.[1]

Amid uncertainty about AI's actual impacts, the effects of regulation, and the costs of inaction, regulatory initiatives generally seek to address the present and future risks commonly associated with AI, including discrimination, violations of privacy and copyright, labour-market disruption, cybercrime, terrorism, and threats to democracy and human existence itself.[2] At the same time, regulation aims not to stifle the technology's revolutionary potential to address major global challenges such as climate change, food security, and disease, in the interest of present and future generations.[3]

In striving to balance innovation and protection, each country selects its own regulatory strategies. The aim is to attract resources and conditions that will enhance competitiveness in the global AI market while finding effective means to protect citizens. This competition, however, can produce the opposite effect. Rather than a "race to the top" with high regulatory standards, it may trigger a "race to the bottom", in which countries loosen their rules to attract investment, compromising safety and protection against AI-related risks.[4]

---

1 SMUHA, Nathalie A. From a "Race to AI" to a "Race to AI Regulation": Regulatory Competition for Artificial Intelligence. *Journal of Law, Innovation and Technology*, v. 13, n. 1, 2021, p. 57–84.

2 ARBEL, Yonathan; TOKSON, Matthew; LIN, Albert. Systemic Regulation of Artificial Intelligence. *Arizona State Law Journal*, Phoenix: Arizona State University, v. 56, n. 2, p. 545–620, Summer 2024.

3 CAJUEIRO, Daniel O.; CELESTINO, Victor R. R. *A Comprehensive Review of Artificial Intelligence Regulation: Weighing Ethical Principles and Innovation*. Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4950727. Accessed: 12 May 2026, p. 12–15.

4 SMUHA, Nathalie A., *op. cit*., p. 60.

Brazil and Australia, though geographically distant and shaped by very different historical trajectories, share notable affinities in geographic, economic, social, and political terms. Both are countries of continental scale, with diverse climates and rich environmental endowments that directly shape their development strategies and international positioning. Both economies depend heavily on natural-resource exports, including agricultural and mineral commodities, and both maintain significant trade relationships with China. Both are multicultural societies marked by ethnic diversity and pluralism, although income distribution differs considerably between them. Both are consolidated democracies, with federal systems, stable institutions, and commitments to the rule of law. They share core liberal-democratic values on market economies, human rights, and the environment, even if engagement levels diverge. Finally, both Brazil and Australia aspire to consolidate themselves as regional powers and have shown interest in cooperating on emerging global agendas.

This article examines Australia's regulatory process for AI in order to draw possible contributions for Brazil's own efforts. The method adopted is a case study, organised in a comparative key, supported by bibliographic and documentary research on Australian initiatives, particularly primary sources from the Australian government, together with academic literature on regulatory theory and AI governance.

The argument develops in four movements. It begins with the conceptual ground (what regulation is, why it is justified, and which strategies are available, especially in the context of emerging technologies). Then it turns to Australia's regulatory trajectory, tracing how an initial commitment to ethical principles, public consultation, and voluntary standards gave way, in late 2025, to an explicit retreat from binding rules in favour of a technology-neutral approach. This trajectory is read against Brazil's own path, where Bill No. 2338/2023 (PL 2338) advances towards a comprehensive risk-based statute, accompanied on the policy and investment side by the Brazilian AI Plan (PBIA 2024–2028). Finally, it closes by drawing mutual lessons, including what Brazil may take from Australia's experience and what Brazil's institutional design may, in turn, offer to the Australian debate.

2. REGULATING AI: RATIONALES AND STRATEGIES

In Philip Selznick's classic formulation, regulation is conceived as sustained and focused control, exercised by a public agency, over activities valued by a community. [5]

Yet the meaning of regulation encompasses a far broader range of instruments and actors. It can be understood as a set of binding legal commands enforced by competent authorities, but also as any deliberate state influence over social and economic behaviour, exercised through subsidies, contracts, or incentives. Beyond the strictly state sphere, regulation may further be understood as any form of influence, including non-governmental influence, that shapes the conduct of individuals or institutions, as occurs through the action of business groups, professional associations, and technical bodies.[6]

The concept of regulation typically evokes the idea of restriction and prohibition. This view, often described as a "red light" approach, frames regulation as an activity that limits behaviour and prevents undesired outcomes. Its focus lies in establishing what cannot be done, imposing barriers and sanctions to deter harmful or unlawful conduct. Regulation, however, is not confined to that role. It may also act as a facilitator. This second posture, the "green light" approach, emphasises regulation as an enabler and structuring force for economic and social activity. Rather than merely barring certain conduct, regulation creates the conditions in which particular activities may unfold in an orderly and productive manner.[7]

These roles are not mutually exclusive. They can overlap, or be pursued simultaneously through different regulatory instruments. In the AI field, the protective ("red light") role aimed at minimising adverse impacts may manifest in transparency obligations or ethical guidelines. The fostering ("green light") role, which seeks to encourage beneficial innovation, may take the form of public funding, tax incentives

---

5 SELZNICK, Philip. Focusing Organisational Research on Regulation. In: NOLL, Roger (ed.). *Regulatory Policy and the Social Sciences*. Berkeley: University of California Press, 1985, p. 363.

6 BALDWIN, Robert; CAVE, Martin; LODGE, Martin. *Understanding Regulation: Theory, Strategy, and Practice*. 2nd ed. Oxford: Oxford University Press, 2012, p. 2–3.

7 *Ibid*., p. 3.

for companies developing AI applications, or policies facilitating the immigration of skilled professionals.[8]

The need for regulation is often justified by the presence of market failures: information asymmetries, negative externalities, natural monopolies, or the absence of competition.[9] Such failures are evident in the AI context. Consumers frequently lack adequate knowledge of how algorithms operate (information asymmetry); the impacts of automated decisions can fall on third parties not involved in the operation (externalities); and the dominance of major digital platforms produces concentration of power and barriers to entry for new actors (monopolistic dynamics).

Regulation can also be grounded in social and ethical considerations, such as the protection of fundamental rights, the promotion of equity, and the strengthening of social solidarity. These are situations in which the state is called upon to intervene, particularly to safeguard vulnerable groups.[10] Here, regulation transcends purely economic logic and assumes a normative-moral function, seeking to correct structural inequalities and prevent abuses. In the AI domain, this ethical dimension is particularly salient. Machine-learning systems can reproduce or amplify biases present in their training data, leading to discrimination on the basis of race, gender, or social class. Moreover, opaque and uninterpretable systems challenge principles such as due process, the duty of transparency in matters of public interest, and accountability, making it difficult to contest automated decisions.[11]

Finally, regulation can be motivated by the need for systemic planning and coordination of a given sector, filling gaps left by private action and preventing normative and institutional fragmentation. As Baldwin, Cave, and Lodge note, where multiple authorities operate in dispersed fashion or with divergent objectives, regulation performs an essential function of rationalisation and integration, responding to the overlap of competences.[12] Coordination, in such cases, may take varied forms, including hierarchical arrangements, collaborative networks, and standardisation

---

8 SMUHA, Nathalie A., op. cit., p. 62–63.

9 BALDWIN, Robert; CAVE, Martin; LODGE, Martin, op. cit., p. 15.

10 *Ibid*., p. 22–23.

11 CAJUEIRO, Daniel O.; CELESTINO, Victor R. R., *op. cit.*, p. 17–23.

12 BALDWIN, Robert; CAVE, Martin; LODGE, Martin, *op. cit.*, p. 366–368.

mechanisms that promote greater predictability and regulatory coherence. This coordinating function is particularly relevant in the case of AI, which cuts across multiple sectors (health, the justice system, public safety, finance, education) and is marked by a high degree of technological uncertainty and rapid, continuous transformation.

Once the rationales for regulation are established, the central question becomes how to regulate. The issue, then, is which strategies or approaches can be employed to achieve the proposed goals, including protection of rights, promotion of equity, encouragement of innovation, and institutional coordination. A wide range of regulatory strategies is available, from traditional command-and-control models to mechanisms based on the market, self-regulation, and responsive regulation.[13] The choice of the most appropriate strategy depends on each country's normative and institutional context, available resources, the regulator's technical capacity, the specific public-policy objectives in view, and the characteristics of the regulated technology or sector.

In the AI field, traditional command-and-control strategies, which involve detailed norms, specific prohibitions, and sanctions for non-compliance, can be effective in restraining conduct that affects fundamental rights, such as discriminatory uses or privacy violations. Yet this model has important limitations, since it tends to be inflexible, reactive, and poorly suited to technological innovation.[14] Other complementary approaches have therefore been proposed and applied.

One such strategy is self-regulation, in which the actors of a given sector (firms, professional associations, technical bodies) establish their own codes of conduct, technical standards, and best practices.[15] In some cases, self-regulation occurs under state oversight, configuring models of co-regulation or meta-regulation, in which the public authority's role is to encourage, supervise, or validate norms produced by private initiative. Suited to the dynamism of technological innovation, this approach can foster an institutional culture of responsibility, although it remains limited in terms of enforcement effectiveness and exposed to the risk of regulatory capture. Its

---

13 Ibid., p. 105–106.

14 CAJUEIRO, Daniel O.; CELESTINO, Victor R. R., *op. cit.*, p. 31–32.

15 BALDWIN, Robert; CAVE, Martin; LODGE, Martin, *op. cit.*, p. 138–157.

success depends, to a great extent, on the alignment between regulated actors' incentives and public objectives, as well as on the existence of effective accountability and monitoring mechanisms.

Another common AI-regulation strategy is risk-based regulation, exemplified by the European Union's AI Act. This approach seeks to calibrate the degree of regulatory intervention according to the gravity and probability of risks associated with different AI systems. Systems considered "high-risk", such as those used in public security, justice, or recruitment, are subjected to more stringent regulatory requirements (impact assessments, transparency obligations, data governance, continuous monitoring), while "low-risk" systems may operate under lighter rules. As Ebers observes, for risk-based regulation to function adequately, the categorisation criteria must be clear, technically grounded, and amenable to updating, given the rapid evolution of the technologies involved.[16] The strategy also demands strong institutional capacity for technical analysis and continuous oversight throughout the AI life cycle, a particular challenge in jurisdictions with fragmented or developing regulatory structures.

In the field of emerging technologies, innovation-oriented regulation also occupies a prominent place. It seeks to reconcile risk control with the promotion of experimentation and technological progress, premised on the view that, in certain sectors, excessive regulatory rigidity may stifle innovation and prevent promising technologies from being developed or safely deployed. To address this dilemma, experimental regulatory environments (sandboxes) are made available, in which companies can develop and test technological solutions under the supervision of regulatory authorities. This format facilitates joint learning between innovators and regulators, allowing for more responsive, evidence-based regulatory adjustments.[17] Its effectiveness, however, demands a delicate balance. The experimental environment cannot become a space of unrestricted permissiveness, and must guarantee minimum protective standards, particularly in respect of fundamental rights. Moreover, the experience accumulated within sandboxes must be systematically incorporated

---

[16] EBERS, Martin. *Truly Risk-Based Regulation of Artificial Intelligence: How to Implement the EU's AI Act*. 2023. Available at: https://ssrn.com/abstract=4496789. Accessed: 12 May 2026.

[17] CAJUEIRO, Daniel O.; CELESTINO, Victor R. R., *op. cit*., p. 34–35.

into broader regulatory frameworks, so that such initiatives do not remain isolated and fail to produce structural impact.

Other relevant approaches include responsive regulation and principles-based regulation. Responsive regulation, proposed by Ayres and Braithwaite, holds that the regulator's response should vary according to the regulated entity's behaviour, beginning with cooperative mechanisms and escalating to coercive measures only when necessary.[18] This approach values dialogue and encourages voluntary compliance, but may be perceived as overly lenient in sectors with high potential for harm. Principles-based regulation, in turn, adopts broad normative guidelines rather than detailed prescriptive rules. This approach is particularly useful in innovation contexts, allowing for greater adaptability, but its effectiveness requires constant interpretation and may produce legal uncertainty, especially where there are power asymmetries or limited regulatory capacity to oversee and enforce the principles in question.[19]

Regulation must therefore be understood as a dynamic, multifaceted process that may take different forms according to the objectives pursued, the risks involved, and the characteristics of the regulated sector. In the AI field, this process becomes even more challenging, given the pace of innovation and the technology's potential social and economic impact. The next section examines Australia's approach, with particular attention to its recent reorientation, in order to extract lessons for Brazil.

3. AUSTRALIA'S AI REGULATORY TRAJECTORY

Having outlined the conceptual foundations, rationales, and main strategies applicable to the regulation of artificial intelligence, this section examines how those theoretical categories have materialised in the Australian experience.

The Australian case is particularly instructive. Between 2019 and 2024, the country built a careful regulatory architecture that combined ethical principles, broad public consultation, voluntary technical standards, and a formal proposal for manda-

---

[18] AYRES, Ian; BRAITHWAITE, John. *Responsive Regulation: Transcending the Deregulation Debate*. Oxford: Oxford University Press, 1992.

[19] BALDWIN, Robert; CAVE, Martin; LODGE, Martin, *op. cit.*, p. 302–310.

tory safeguards in high-risk settings. In late 2025, however, that trajectory was sharply redirected. The voluntary standard was superseded by a new Guidance for AI Adoption, an AI Safety Institute was announced as an advisory body, and the National AI Plan formally abandoned the move towards binding rules in favour of a technology-neutral approach grounded in pre-existing laws. The Australian experience thus offers a case study not of linear progression from soft to hard law, but of regulatory ambition followed by deliberate retrenchment.

The analysis is organised around four moments, namely the ethical principles adopted in 2019, the public consultation conducted from 2023 to 2024, the voluntary safety standard of 2024 and its replacement in 2025, and the proposed mandatory guardrails together with the late-2025 pivot away from them. Together, these moments allow for an assessment of how different regulatory strategies were combined in practice, and offer significant material for reflection on the future of AI regulation in Brazil.

3.1 Ethical principles as soft-law foundations

The formulation of ethical principles for artificial intelligence in Australia represents a milestone in the country's efforts towards responsible technological governance, oriented to reconciling innovation with the protection of fundamental rights. These principles were developed by the Australian government's Department of Industry, Science and Resources (DISR) to provide guiding directives for the design, development, and use of AI systems, in both the public and private sectors. They constituted an initial normative response, voluntary and non-binding in character.

The drafting process involved broad public participation and drew on international models of ethical governance. In 2019, a preliminary document with an initial proposal of ethical principles, developed by a specialised division of Australia's national science agency (CSIRO Data61), was submitted to public consultation, attracting more than 130 contributions from industry, academia, government bodies, and civil-society representatives.[20] In addition to those contributions and the

[20] AUSTRALIA, Department of Industry, Science and Resources. *Developing the AI Ethics Principles*. Available at: https://www.industry.gov.au/publications/australias-artificial-intelligence-

participation of expert groups, the principles drew on the main international references published that same year, including the OECD Principles on AI[21] and the European Commission's Ethics Guidelines for Trustworthy AI.[22]

The result of this process was the publication of Australia's AI Ethics Principles, a set of eight principles aimed at ensuring that AI systems operate in a manner that is fair, safe, accountable, and consistent with democratic values. Although they have no legally binding force, the principles are recognised as relevant regulatory instruments for shaping good practices and as a foundation for future formal regulation.[23]

The first principle establishes that the development and use of AI must benefit people, society, and the environment. This is a general directive, requiring that the potential social and environmental impacts of technological applications be considered, assessed, and addressed from the earliest project phases. The notion of "well-being" here encompasses not only economic gains but also the improvement of living conditions, social inclusion, and environmental sustainability.

The centrality of the human person, as the second principle, holds that AI systems must be designed to respect human rights, diversity, and individual autonomy. Automated decisions must not unduly displace human judgement, nor should they manipulate user choices in opaque or coercive ways. In practical terms, this principle calls for explanatory interfaces, mechanisms of meaningful human supervision, and guarantees of informed consent in situations where automated systems directly affect persons.

Fairness comes next, conveying the promotion of equity and the prevention of discrimination. Its implementation requires that AI systems be designed and operated in ways that do not reproduce or intensify existing social inequalities, demanding

ethics-principles/developing-ai- ethics -principles. Accessed: 12 May 2026.

21 OECD. *AI Principles*. Available at: https://www.oecd.org/en/topics/ai-principles.html. Accessed: 12 May 2026.

22 EUROPEAN COMMISSION. *Ethics Guidelines for Trustworthy AI*. Available at: https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai. Accessed: 12 May 2026.

23 AUSTRALIA, Department of Industry, Science and Resources. *Australia's AI Ethics Principles*. Available at: https://www.industry.gov.au/publications/australias-artificial-intelligence-ethics-principles/australias-ai-ethics-principles. Accessed: 12 May 2026.

attention to the quality, representativeness, and integrity of the data used in training. In practice, observance of this principle calls for algorithmic audits, bias testing, and continuous review of decision-making models, particularly in critical contexts such as access to healthcare, credit decisions, recruitment processes, and the operation of justice and public-security systems.

Privacy protection and data security form the fourth principle. It establishes that the development and use of AI systems must respect users' rights, ensuring that the collection, storage, and processing of data occurs securely, transparently, and in accordance with informational self-determination. To this end, recommended practices include data minimisation, the use of anonymisation techniques, the application of robust security protocols, and clear communication regarding the terms of data use and sharing.

Transparency and explainability, in turn, address the intelligibility of automated systems. The aim is that decision-making processes carried out by AI be understandable to users and amenable to audit. In high-impact applications, this translates into adequate documentation of models, the provision of comprehensible justifications for AI-generated decisions, and a preference for interpretable algorithmic models. The principle is particularly relevant where decisional opacity compromises fundamental rights, including the right to information and the possibility of contestation. It is also essential to ensure that people can identify when they are interacting with an AI system.

The sixth principle concerns technical robustness and safety. AI systems must be technically resilient and operate as expected, even in the face of failures, external attacks, or adverse conditions. This calls for rigorous testing and validation, as well as continuous monitoring throughout the technology's life cycle. In contexts of significant uncertainty, the principle aligns with the precautionary logic, recommending a prudent regulatory posture in the face of imperfectly understood risks.

Accountability, the seventh principle, highlights the need for clear mechanisms of duty-attribution and responsibility for harms caused by AI systems, allowing for the identification of developers, providers, and institutional users. It encourages the adoption of internal governance instruments, compliance programmes, and structured channels for accountability.

The eighth and final principle, contestability, ensures that individuals affected by automated systems have the right to challenge decisions that significantly affect them. This right of contestation is essential for the protection of individual safeguards and presupposes the existence of accessible complaint channels, effective review mechanisms, and, where applicable, forms of redress.

To verify the practical applicability of these principles, the Australian government conducted, between 2020 and 2021, a pilot project with public and private organisations, testing the integration of the ethical principles into their workflows and AI systems. The results showed that, although the principles enjoyed broad acceptance, their operationalisation required investment in internal capacity-building, the development of specific tools, and adaptation to different sectoral contexts.[24] The pilot also revealed that, given the more general, abstract, and at times indeterminate character of the ethical principles, their effective implementation required tools of operational translation, including technical guides, performance indicators, and examples of good practice.

In response, in June 2023, the Australian government published a practical guide offering resources to assist organisations in applying the principles at different stages of AI system development.[25] This material presents twenty-six suggested practices for the concrete implementation of the ethical principles, encompassing internal policies, impact-assessment instruments, software tools, and oversight mechanisms.

In addition to the practical guide, with a view to coordinated application of the ethical principles across different levels of Australian public administration, the National Framework for the Assurance of Artificial Intelligence in Government was launched in June 2024, establishing directives for governance, data management, risk assessment, alignment with standards, and ethics in procurement.[26] As a comple-

---

[24] AUSTRALIA, Department of Industry, Science and Resources. *Testing the AI Ethics Principles*. Available at: https://www.industry.gov.au/publications/australias-artificial-intelligence-ethics-framework/testing-ai-ethics-principles. Accessed: 12 May 2026.

[25] REID, Alistair; O'CALLAGHAN, Simon; LU, Yaya. *Implementing Australia's AI Ethics Principles: A Selection of Responsible AI Practices and Resources*. Sydney: Gradient Institute; CSIRO, 2023. Available at: https://www.industry.gov.au/publications/implementing-australias-ai-ethics-principles-selection-responsible-ai-practices-and-resources. Accessed: 12 May 2026.

[26] AUSTRALIA, Department of Finance. *National Framework for the Assurance of Artificial Intelli-*

ment, the Policy for the responsible use of AI in government, in force since September 2024, set obligations for federal agencies and bodies, including the designation of internal officers, the implementation of training programmes, and the publication of transparency statements.[27]

The Australian experience in formulating and disseminating ethical principles for AI thus reflects a systematic effort to anticipate risks and orient technological development responsibly. Although the principles are not binding, their influence has been significant in shaping a regulatory culture grounded in democratic values and in encouraging responsible self-regulation. This ethical foundation would be tested and expanded in the public-consultation process that opened in 2023, examined in the next subsection.

3.2 Consultation, risk, and the case for mandatory safeguards

In June 2023, Australia's Department of Industry, Science and Resources (DISR) launched a public consultation entitled Safe and Responsible AI in Australia, with the aim of fostering a national debate on the opportunities and risks associated with the development and use of AI systems. The consultation process was structured to gather contributions from a wide range of social actors, including industry, academia, government agencies, civil-society organisations, and citizens, in order to guide possible regulatory, ethical, and technical responses to the country's growing adoption of AI.

The discussion paper outlined the national and international landscape of AI governance, described a wide range of regulatory strategies that could be deployed, and raised critical questions on regulatory gaps, guiding principles, and measures to enhance public trust in automated systems.[28]

---

gence in Government. June 2024. Available at: https://www.finance.gov.au/government/public-data/data-and-digital-ministers-meeting/national-framework-assurance-artificial-intelligence-government. Accessed: 12 May 2026.

27 AUSTRALIA, Digital Transformation Agency. *Policy for the Responsible Use of AI in Government*. Available at: https://www.digital.gov.au/policy/ai/policy. Accessed: 12 May 2026.

28 AUSTRALIA, Department of Industry, Science and Resources. *Safe and Responsible AI in Australia: Discussion Paper*. June 2023. Available at: https://storage.googleapis.com/converlens-au-

The contributions received in the consultation revealed substantial convergence around the social, legal, and technical risks associated with AI use, particularly in high-impact contexts. Recurrent concerns included the risk of discriminatory automated decisions, the opacity of algorithmic systems, the fragility and lack of transparency of training data, the difficulty of attributing legal responsibility when systems fail, and the absence of effective mechanisms of contestation and redress. Particular concern was directed at AI foundation models, including generative language and image systems, whose scale, complexity, and cross-cutting use posed heightened risks that general-purpose or sector-specific regulatory frameworks were ill-equipped to address.[29]

Many submissions argued that the predominance of voluntary measures and generic ethical principles, while valuable as a starting point, was proving insufficient to confront these challenges, and ought to be complemented by binding legal instruments capable of protecting fundamental rights and ensuring accountability. The contributions also stressed the need for a coherent national regulatory strategy, capable of avoiding normative fragmentation and promoting integration across different levels and bodies of government, with alignment to emerging international standards (notably those proposed by the European Union and the OECD), so that Australian citizens might enjoy a level of protection equivalent to that offered in other advanced liberal democracies.

In response to the consultation process, in January 2024 the Australian government published a document entitled Safe and Responsible AI in Australia: Interim Response, which consolidated the main findings of the consultation phase and presented a series of provisional directives for the future regulation of AI.[30] The interim response confirmed the government's intention to continue promoting multistake-

industry/industry/p/prj2452c8e24d7a400c72429/public_assets/Safe-and-responsible-AI-in-Australia-discussion-paper.pdf. Accessed: 12 May 2026.

29 AUSTRALIA, Department of Industry, Science and Resources. *Safe and Responsible AI in Australia Consultation: Australian Government's Interim Response*. January 2024. Available at: https://storage.googleapis.com/converlens-au-industry/industry/p/prj2452c8e24d7a400c72429/public_assets/safe-and-responsible-ai-in-australia-governments-interim-response.pdf. Accessed: 12 May 2026.

30 *Ibid*.

holder dialogue, and proposed a risk-based regulatory approach with a view to ensuring that AI systems deployed in legitimate but high-risk contexts (such as healthcare, criminal justice, finance, or education) would be implemented in a safe, auditable, and accountable manner.

The interim response observed that the legal frameworks in force were not sufficient to address the specific risks arising from AI, particularly in matters of algorithmic discrimination, accountability, and transparency. Accordingly, in addition to the review of certain existing laws (privacy, anti-discrimination, consumer protection, competition, intellectual property, and product safety), the government proposed the introduction of mandatory safeguards, especially for high-risk AI applications, contemplating obligations such as impact testing, traceability, disclosure of system purpose, and labelling of AI-generated content.

The interim response also stressed the role of voluntary initiatives, suggesting the use of technical standards, codes of conduct, and regulatory sandboxes as complementary forms of governance. The document further reaffirmed the government's commitment to strengthening public trust in AI through transparency measures, accessible mechanisms of contestation, and digital-education initiatives, and underscored the importance of aligning future Australian regulation with international norms and guidelines, particularly those produced by the OECD, the European Union, and multilateral fora dedicated to AI safety.

While not itself a normative instrument, the interim response signalled the direction the Australian government intended to follow, namely a modular regulation, calibrated according to risk, open to innovation, and centred on the protection of fundamental rights. That direction converged with the hybrid regulatory models then under international debate, in which mandatory norms coexist with soft-law instruments, technical standards, and incentives to self-regulation. Two parallel instruments would emerge from this design in September 2024, namely a voluntary safety standard and a proposals paper introducing mandatory guardrails for high-risk settings, which are examined in the following subsections.

3.3 Voluntary guardrails: from VAISS to AI6

In September 2024, in line with the direction announced in the interim response,

the Australian government released the Voluntary AI Safety Standard (VAISS), developed in consultation with technical experts, industry representatives, and academics. The instrument set out a uniform set of recommended practices to guide organisations developing or deploying AI in Australia, with particular attention to those operating in sectors of greater social or economic sensitivity.[31]

Drawing on technical references such as the United States NIST AI Risk Management Framework and ISO/IEC 42001:2023 on AI management systems, the VAISS proposed ten fundamental safeguards, or guardrails, designed to structure an organisational culture of responsibility, technical integrity, fairness, and accountability throughout the development and use of AI. The instrument was explicitly aligned with a risk-based and human-centred approach, in keeping with the Bletchley Declaration adopted at the November 2023 AI Safety Summit.[32]

The ten guardrails covered, in summary form, the principal dimensions of AI governance now consolidated in international practice. They required organisations to establish clear processes of internal accountability and governance for AI projects; to undertake proactive risk assessment across the system life cycle; to ensure data governance and the technical security of AI systems; to conduct rigorous performance testing and continuous post-deployment monitoring; to maintain meaningful human oversight, including the possibility of human review, contestation, and reversal of critical automated decisions, in line with the human-in-the-loop logic defended by various international authors and bodies;[33] to inform end users adequately about the nature and operation of AI systems with which they interact; to provide accessible mechanisms for affected individuals to challenge automated decisions; to ensure transparency along the AI supply chain, sharing relevant information with

---

31 AUSTRALIA, Department of Industry, Science and Resources. *Voluntary AI Safety Standard*. September 2024. Available at: https://www.industry.gov.au/publications/voluntary-ai-safety-standard. Accessed: 12 May 2026.

32 UNITED KINGDOM. *The Bletchley Declaration by Countries Attending the AI Safety Summit, 1–2 November 2023*. London: UK Government, 2023. Available at: https://www.gov.uk/government/publications/ai-safety-summit-2023-the-bletchley-declaration/the-bletchley-declaration-by-countries-attending-the-ai-safety-summit-1-2-november-2023. Accessed: 12 May 2026.

33 PARENTONI, Leonardo. What Should We Reasonably Expect from Artificial Intelligence? *SSRN Electronic Journal*, v. XII, n. 02, p. 179–219, July 2022.

partners, providers, and users; to maintain robust organisational records allowing third-party verification of compliance; and, finally, to engage stakeholders in the design and use of AI systems, with particular attention to safety, diversity, inclusion, and equity.

Although adherence to the VAISS was voluntary, the instrument formed part of a broader movement of responsive regulation and soft law, intended to prepare the Australian technological ecosystem for a possible mandatory regime, particularly in respect of AI systems used in high-risk contexts. The clarity and technical breadth of the ten guardrails were designed to provide practical directives that could be incorporated by different sectors, while at the same time paving the way for future legal obligations.

Nevertheless, that trajectory shifted in October 2025. The National AI Centre (NAIC), within the Department of Industry, Science and Resources, replaced the VAISS with a new instrument, the Guidance for AI Adoption, which condensed the ten guardrails into six essential practices, often referred to in shorthand as "AI6".[34] The new guidance was framed as a response to rapid developments in AI technology and governance over the preceding twelve months, and to feedback received from industry on the operationalisation of the VAISS.

The six essential practices comprised deciding who is accountable for each AI use case; understanding impacts on people, communities, and the environment; measuring and managing risks across the system life cycle; sharing information with the appropriate stakeholders, internal and external; testing and monitoring AI systems before and after deployment; and maintaining meaningful human oversight throughout. In substance, the AI6 retained the conceptual core of the ten VAISS guardrails, but recast it in a more operationally accessible form, with two complementary versions of the guidance, namely a Foundations document for organisations beginning their AI journey, and an Implementation Practices document for governance professionals and technical experts.

Read in isolation, the move from VAISS to AI6 might appear as a modest techni-

[34] AUSTRALIA, Department of Industry, Science and Resources; National AI Centre. *Guidance for AI Adoption*. October 2025. Available at: https://www.industry.gov.au/publications/guidance-for-ai-adoption. Accessed: 12 May 2026.

cal consolidation. Read against what was happening in parallel in the debate on mandatory rules, however, the substitution acquires a different significance. By the time AI6 was released, the policy orientation towards mandatory guardrails had already begun to shift, and AI6 effectively repositioned voluntary guidance as the principal reference for AI governance in Australia, rather than as a preparatory stage for binding regulation. The next subsection traces that shift in detail.

3.4 Mandatory guardrails and the technology-neutral pivot

In parallel with the launch of the VAISS in September 2024, and recognising that voluntary measures, however well structured, would be insufficient to address the systemic risks and negative externalities potentially associated with AI in high-risk contexts, the Australian government published the Proposals Paper for Introducing Mandatory Guardrails for AI in High-Risk Settings.[35] The paper was a preliminary proposal, drawn from the recommendations of an Expert Group convened in the wake of the interim response, and was intended to guide dialogue between government, civil society, the productive sector, and academia on the design of binding rules for high-risk applications.

The proposal departed from the recognition that, as observed in the International Scientific Report on the Safety of Advanced AI, AI presents characteristics that distinguish it from other technologies, namely degrees of autonomy that bypass human intervention, general cognitive capacities, ubiquity, versatility, capacity for learning and adaptation, exponential gains in speed and scale, opacity and limited explainability, and the ability to emulate human behaviour with notable realism.[36] The use of such systems in critical sectors (surveillance, predictive policing, credit, labour rela-

---

[35] AUSTRALIA, Department of Industry, Science and Resources. *Safe and Responsible AI in Australia: Proposals Paper for Introducing Mandatory Guardrails for AI in High-Risk Settings*. September 2024. Available at: https://storage.googleapis.com/converlens-au-industry/industry/p/prj2f6f02ebfe6a8190c7bdc/page/proposals_paper_for_introducing_mandatory_guardrails_for_ai_in_high_risk_settings.pdf. Accessed: 12 May 2026.

[36] UNITED KINGDOM. *International Scientific Report on the Safety of Advanced AI: Interim Report*. London: GOV.UK, 2024. Available at: https://www.gov.uk/government/publications/international-scientific-report-on-the-safety-of-advanced-ai. Accessed: 12 May 2026.

tions, healthcare, education) could produce significant adverse effects on individuals, communities, and democratic institutions. According to the proposal, such systems could amplify inequalities, reinforce historical biases, or compromise fundamental rights, in addition to being susceptible to use in the commission of crimes and disinformation campaigns, with risks to national security and democracy. There were also risks associated with emergent behaviours and the potential loss of control over AI agents.

The proposal accordingly advanced a risk-based regulatory model, through the adoption of mandatory rules specific to high-risk situations, with a focus on preventive measures (*ex ante* and pre-market), distributed across the AI life cycle (design, data collection and preparation, training and modelling, evaluation, deployment, operation, and monitoring) and across the entire AI supply chain, encompassing both developers and deployers.

The document presented for debate a definition of high-risk AI structured along two axes. The first considered applications whose use is known or foreseeable, weighing the gravity and probability of adverse impacts. To assess the degree of risk of each application and its classification as high-risk, the proposal offered guiding principles considering factors such as violations of human rights, harms to physical and mental health and safety, significant adverse legal effects, collective and systemic impacts on the economy, society, the environment, and the Australian legal order, as well as the gravity and extent of those harms. The document provided examples of how to apply those principles in assessing risk, presented a table of high-risk AI use cases identified in other jurisdictions, and invited contributions on low-risk situations that might be affected by regulation built around a defined list of high-risk cases. It also sought input on AI use cases of unacceptable risk that ought to be prohibited.

The second axis covered general-purpose AI models (GPAI), whose versatility and scale make their risks difficult to anticipate. Here, the proposal followed an approach similar to that of Canada, both as regards the broad definition of GPAI ("an AI model capable of being used, or adapted for use, for a variety of purposes, whether for direct use or for integration into other systems") and the application of mandatory safeguards to any model meeting that definition, on the understanding that high risk is inherent to such models. The document also requested suggestions on how the mandatory safeguards would apply to general-purpose models, and whether differ-

ent or supplementary safeguards would be needed, as contemplated by the European Union's AI Act for GPAI presenting systemic risk.

To mitigate the risks associated with AI, the document proposed ten mandatory safeguards for developers and deployers of AI systems in high-risk contexts, including general-purpose models, applicable to each stage of the system life cycle. The proposed safeguards drew on the European Union's AI Act, the Canadian proposal for an Artificial Intelligence and Data Act (AIDA), and various technical standards. [37] Their substance covered clear processes of accountability and governance, risk-management mechanisms, data governance and security, ongoing testing and monitoring, meaningful human oversight, transparency and explainability for users, mechanisms of contestation for affected persons, information sharing along the supply chain, record-keeping and technical documentation enabling compliance assessment, and the conduct of conformity assessments. With the exception of the last item, the safeguards aligned with those set out in the VAISS.

Three regulatory pathways were presented for the implementation of the mandatory safeguards: (i) the adaptation of existing sectoral legal frameworks to incorporate the safeguards; (ii) the adoption of framework legislation containing definitions, thresholds, and safeguards, complemented by amendments to existing laws to confer powers on sectoral authorities; or (iii) the creation of a comprehensive AI Act, applicable across sectors, including definitions, thresholds, safeguards, and enforcement and sanction mechanisms.

In November 2024, one month after the publication of the Proposals Paper, the Australian Senate Select Committee on Adopting Artificial Intelligence delivered its final report, comprising thirteen recommendations.[38] The first of those recommen-

---

[37] ISO/IEC 42001:2023 (Artificial Intelligence Management System); ISO/IEC 23894:2023 (Information Technology — Artificial Intelligence — Guidance on Risk Management); ISO/IEC TR 29119-11:2020 (Software and Systems Engineering — Software Testing — Part 11: Guidelines on the Testing of AI-Based Systems); ISO/IEC 24027:2022 (Information Technology — Artificial Intelligence — Bias in AI Systems and AI Aided Decision Making).

[38] AUSTRALIA, Parliament. Senate. Select Committee on Adopting Artificial Intelligence (AI). *Adopting Artificial Intelligence (AI): Report*. Canberra: Parliament of Australia, November 2024. Available at: https://www.aph.gov.au/Parliamentary_Business/Committees/Senate/Adopting_Artificial_Intelligence_AI/AdoptingAI/Report. Accessed: 12 May 2026.

dations was precisely the introduction of a specific and comprehensive law to regulate high-risk AI uses, in line with option (iii) of the Proposals Paper. The Committee also recommended a principles-based approach for the definition of high-risk AI, supplemented by a non-exhaustive list of expressly defined use cases (Recommendation 2), including general-purpose models such as large language models (LLMs) (Recommendation 3). The remaining recommendations addressed varied questions, including measures to confront the impacts of AI on the labour market and on copyright.

Although the trajectory traced above pointed in the direction of binding rules, the Australian regulatory orientation shifted markedly during 2025. The pivot did not occur abruptly. It crystallised through a sequence of institutional acts that, taken together, redirected the path established between 2023 and 2024.

The first signal came in August 2025, when the Productivity Commission, an independent advisory body to the federal government on micro-economic policy, released an interim report on data and digital technology recommending that AI-specific regulation be adopted only as a last resort, that all work on the proposed economy-wide mandatory guardrails for high-risk AI be paused, and that priority be given to a gap analysis of existing technology-neutral frameworks before any AI-specific intervention.[39] The report framed AI primarily as a productivity opportunity, and warned that AI-specific rules built on top of existing frameworks risked producing duplication, inefficiency, and a drag on innovation. The recommendation found resonance in submissions from the technology sector, which had consistently argued that regulatory responses should build on existing technology-neutral instruments rather than introduce a new AI-specific regime.

The reorientation was institutionalised in late 2025 through two further acts. In November 2025, the government announced the establishment of an Australian AI Safety Institute (AISI), to become operational in early 2026, with the mandate to monitor, test, and share information on emerging AI capabilities, risks, and harms,

[39] PRODUCTIVITY COMMISSION. *Harnessing Data and Digital Technology: Interim Report*. Canberra: Productivity Commission, August 2025. Available at: https://www.pc.gov.au/inquiries-and-research/data-digital/interim/. Accessed: 17 May 2026.

and to advise government on areas of regulatory reform and capability gaps.[40] Crucially, the AISI was designed as an advisory body, working alongside the existing specialist regulators (notably the Office of the Australian Information Commissioner for privacy, the Australian Competition and Consumer Commission for consumer protection, and the eSafety Commissioner for online harms), but without enforcement powers of its own.

The decisive act came on 2 December 2025, with the publication of the National AI Plan, structured around three goals, namely capturing AI opportunities, spreading the benefits, and keeping Australians safe.[41] On the regulatory dimension, the Plan formally abandoned the proposal for mandatory guardrails in high-risk settings, in favour of a two-pronged approach. The first prong is the uplift and clarification of existing technology-neutral laws (privacy, consumer protection, anti-discrimination, online safety, sectoral regulation), with targeted amendments where emerging risks require them. The second is the issuance of further voluntary guidance, of which the Guidance for AI Adoption already constituted the leading example, complemented by the operations of the AISI. The Plan thus reframed the regulatory question. Rather than designing a binding regime tailored to AI, the government reaffirmed the technology-neutral character of the pre-existing legal framework as fit-for-purpose, requiring only targeted adjustments rather than wholesale reform.

This pivot was not uncontested. Critics observed that technology-neutral instruments may prove ill-suited to address AI-specific harms, including those associated with deepfakes, algorithmic bias, and autonomous decision-making. They further argued that the abandonment of binding rules placed Australia at a more permissive end of the international regulatory spectrum, whereas peer jurisdictions, the European Union foremost among them, retained AI-specific obligations as the regulatory baseline.[42]

---

[40] AUSTRALIA, Department of Industry, Science and Resources. *Australia to Establish New Institute to Strengthen AI Safety*. Press release, 25 November 2025. Available at: https://www.industry.gov.au/news/australia-establish-new-institute-strengthen-ai-safety. Accessed: 12 May 2026.

[41] AUSTRALIA, Department of Industry, Science and Resources. *National AI Plan 2025*. Canberra: DISR, 2 December 2025. Available at: https://www.industry.gov.au/sites/default/files/2025-12/national-ai-plan.pdf. Accessed: 12 May 2026.

[42] For academic critiques of the pivot from binding rules to a technology-neutral approach, see, in

These criticisms carry weight, particularly when measured against the European Union as the foremost peer benchmark. Read against the wider international landscape, however, Australia's pivot also appears consistent with a recent broader recalibration of AI regulation across comparable jurisdictions. In the European Union, although the AI Act remained the central regulatory baseline, the Digital Omnibus proposal introduced a simplification agenda directed at easing implementation, including adjustments to the application of high-risk AI rules and simplified technical documentation requirements for smaller firms.[43] The United Kingdom continued to favour a pro-innovation, regulator-led model over a comprehensive AI-specific statute,[44] and the United States moved away from the Biden administration's safety-oriented federal framework after the revocation of Executive Order 14110.[45] On this

particular, GOLDENFEIN, Jake; PARKER, Christine; WEATHERALL, Kimberlee. Australia's National AI Plan Has Just Been Released: Who Exactly Will Benefit? *The Conversation*, 3 December 2025, available at: https://theconversation.com/australias-national-ai-plan-has-just-been-released-who-exactly-will-benefit-270981, accessed 17 May 2026; and PANE, John (Electronic Frontiers Australia), cited in IIS Partners, *Australia's National AI Plan: Big Vision, Missing Guardrails*, 15 December 2025, available at: https://www.iispartners.com/insights/2025/12/15/australias-national-ai-plan-big-vision-missing-guardrails, accessed 17 May 2026. For a multi-author expert reaction collected by the Australian Science Media Centre, see SCIMEX. *Expert Reaction: National AI Plan Shifts Away from Mandatory Guardrails*. 2 December 2025. Available at: https://www.scimex.org/newsfeed/expert-reaction-national-ai-plan-shifts-away-from-mandatory-guardrails. Accessed: 17 May 2026.

43 EUROPEAN COMMISSION. *Digital Omnibus on AI Regulation: Proposal for a Regulation amending Regulations (EU) 2024/1689 and (EU) 2018/1139 as regards the simplification of the implementation of harmonised rules on artificial intelligence*, COM(2025) 836 final, 2025/0359 (COD). Brussels, 19 November 2025. Available at: https://digital-strategy.ec.europa.eu/en/library/digital-omnibus-ai-regulation-proposal. Accessed: 18 May 2026.

44 UNITED KINGDOM, Department for Science, Innovation and Technology. *A pro-innovation approach to AI regulation*. White Paper, CP 815. London: DSIT, March 2023. Available at: https://www.gov.uk/government/publications/ai-regulation-a-pro-innovation-approach. Accessed: 18 May 2026. See also the subsequent UNITED KINGDOM, Department for Science, Innovation and Technology. *A pro-innovation approach to AI regulation: government response to consultation*. London: DSIT, February 2024. Available at: https://www.gov.uk/government/consultations/ai-regulation-a-pro-innovation-approach-policy-proposals/outcome/a-pro-innovation-approach-to-ai-regulation-government-response. Accessed: 18 May 2026.

45 UNITED STATES. *Executive Order 14179 of January 23, 2025*, Removing Barriers to American

reading, Australia's shift could be presented not as a retreat from peer practice, but as an attempt to align itself with a wider movement among advanced liberal democracies towards more flexible, innovation-sensitive forms of AI governance.

In summary, the Australian trajectory between 2019 and 2025, taken as a whole, offers a cautionary case study rather than a model of progressive escalation. A regulatory ambition carefully prepared over five years, supported by extensive consultation, expert input, and detailed technical groundwork, ultimately did not translate into a binding regime. What persists, in the post-2025 landscape, is a hybrid arrangement in which AI-specific instruments operate either as voluntary guidance (the Guidance for AI Adoption and its underlying Ethics Principles) or as institutional and advisory capacity (the AISI), while binding obligations are channelled through pre-existing technology-neutral statutes, supplemented by targeted reforms such as the forthcoming Privacy Act amendments on automated decision-making transparency. Whether this arrangement proves adequate to address AI-specific harms remains an open question. For jurisdictions such as Brazil, still shaping AI regulatory frameworks built around binding obligations, the Australian case therefore provides a useful comparative counterpoint.

4. AUSTRALIA AND BRAZIL IN COMPARATIVE PERSPECTIVE: DIVERGENT PATHS, MUTUAL LESSONS

The trajectory examined above, when read against Brazil's contemporaneous path, reveals not a relation of model and follower but a comparison of two democracies that have taken substantially different roads in the same regulatory landscape. The interest of the comparison lies precisely in the divergence. Where Australia ultimately retreated from binding rules, Brazil has continued, however slowly, to advance towards a comprehensive risk-based statute. Each path carries lessons that the other may consider.

Brazil's engagement with AI regulation builds on a digital-rights tradition that

Leadership in Artificial Intelligence. Federal Register, v. 90, n. 20, p. 8741, 31 January 2025. Available at: https://www.federalregister.gov/documents/2025/01/31/2025-02172/removing-barriers-to-american-leadership-in-artificial-intelligence. Accessed: 18 May 2026.

long predates the present moment. The Marco Civil da Internet of 2014 established the foundational principles for internet governance,[46] and the Lei Geral de Proteção de Dados (LGPD) of 2018 introduced a general data protection regime broadly comparable in scope to the European GDPR,[47] with the Agência Nacional de Proteção de Dados (ANPD), Brazil's data protection authority, as its central enforcement body. In April 2021, the federal government published the Estratégia Brasileira de Inteligência Artificial (EBIA), a strategic document setting out priorities for AI research, application, and governance.[48] That groundwork is significant. When Brazil's legislative process on AI began in earnest, it was building on an existing institutional infrastructure, with an established data protection authority, a recognised framework of digital rights, and a constitutional commitment to the dignity of the human person that would explicitly anchor the AI bill itself.

Brazil's principal regulatory instrument, Bill No. 2338/2023 (PL 2338), was introduced in the Federal Senate in May 2023.[49] The bill drew substantially from the work of a Commission of Jurists convened by the Senate in 2022, whose final report had proposed a comprehensive risk-based framework explicitly inspired by the European Union's AI Act, while incorporating elements of the Brazilian constitutional tradition. After eighteen months of work, including public hearings, sectoral consultations, and several rounds of substantive amendment, PL 2338 was approved by the Senate in December 2024 and forwarded to the Chamber of Deputies.

---

46 BRASIL. Lei nº 12.965, de 23 de abril de 2014. Estabelece princípios, garantias, direitos e deveres para o uso da Internet no Brasil (Marco Civil da Internet). Available at: https://www.planalto.gov.br/ccivil_03/_ato2011-2014/2014/lei/l12965.htm. Accessed: 12 May 2026.

47 BRASIL. Lei nº 13.709, de 14 de agosto de 2018. Lei Geral de Proteção de Dados Pessoais (LGPD). Available at: https://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/l13709.htm. Accessed: 12 May 2026.

48 BRASIL, Ministério da Ciência, Tecnologia e Inovações. *Estratégia Brasileira de Inteligência Artificial (EBIA)*. Brasília: MCTI, 2021. Available at: https://www.gov.br/mcti/pt-br/acompanhe-o-mcti/transformacaodigital/inteligencia-artificial. Accessed: 12 May 2026.

49 BRASIL, Senado Federal. Projeto de Lei nº 2338, de 2023. Dispõe sobre o uso da inteligência artificial. Available at: https://www25.senado.leg.br/web/atividade/materias/-/materia/157233. Accessed: 12 May 2026. For the report of the Commission of Jurists that informed the bill, see SENADO FEDERAL. *Comissão de Juristas Responsável por Subsidiar Elaboração de Substitutivo sobre Inteligência Artificial no Brasil: Relatório Final*. Brasília: Senado Federal, December 2022.

In the Chamber, a special committee was established in March 2025 to examine the bill, with broad cross-party composition and a work plan structured around extensive participatory consultation. Between May and September 2025, the committee held twelve public hearings, complemented by regional seminars across all five regions of the country and an international seminar,[50] in an exercise of public consultation comparable in scope to the Australian process of 2023 and arguably more institutionally embedded, given its location within the legislative procedure itself. The hearings drew contributions from industry, academia, civil-society organisations, regulatory bodies, the judiciary, and the federal prosecutor's office, generating an extensive documentary record that has informed the drafting of the rapporteur's report.

In December 2025, the Executive Branch supplemented this trajectory by submitting a second bill (PL 6237/2025), designed to remedy a constitutional defect of initiative identified in PL 2338, which had attributed competences to the ANPD, a matter within the Executive's exclusive initiative under Brazilian constitutional law. The second bill articulates the institutional architecture of what the original bill had called the Sistema Nacional de Regulação e Governança de Inteligência Artificial (National System for AI Regulation and Governance, SIA).[51] Under the proposed design, the SIA coordinates the regulatory, supervisory, and sanctioning activities of multiple bodies, with the ANPD serving as the central coordinating authority re-

---

[50] CÂMARA DOS DEPUTADOS. Comissão Especial destinada a proferir parecer ao Projeto de Lei nº 2338, de 2023, sobre Inteligência Artificial. *Plano de Trabalho do Relator and related materials*. Available at: https://www2.camara.leg.br/atividade-legislativa/comissoes/comissoes-temporarias/especiais/57a-legislatura/comissao-especial-sobre-inteligencia-artificial-pl-2338-23. Accessed: 12 May 2026. For a systematisation of the contributions to the public hearings, see DATA PRIVACY BRASIL RESEARCH. *Audiências Públicas da Comissão Especial de Inteligência Artificial da Câmara dos Deputados: Dinâmicas Participativas e Temas Emergentes em 2025*. São Paulo, October 2025. Available at: https://www.dataprivacybr.org/relatorio-audiencias-publicas-da-comissao-especial-de-inteligencia-artificial-da-camara-dos-deputados-dinamicas-participativas-e-temas-emergentes-em-2025/. Accessed: 18 May 2026.

[51] BRASIL, Câmara dos Deputados. Projeto de Lei nº 6237, de 2025. Institui o Sistema Nacional para Desenvolvimento, Regulação e Governança de Inteligência Artificial — SIA. Brasília, 8 December 2025. Available at: https://www.camara.leg.br/proposicoesWeb/fichadetramitacao?idProposicao=2594000. Accessed: 12 May 2026.

sponsible for issuing general norms and for regulating sectors not covered by an existing sectoral regulator. Sectoral authorities retain their regulatory powers within their respective domains, with the ANPD operating in a residual and harmonising capacity. The bill also creates a strategic policy body, named the Conselho Brasileiro de Inteligência Artificial (CBIA), composed of up to five ministries together with the ANPD, alongside two consultative committees representing civil society, the productive sector, and the scientific community. As of mid-2026, the substantive bill (PL 2338) remains pending the rapporteur's formal report in the Chamber's special committee, with prospects of approval affected by Brazil's electoral cycle. The institutional bill (PL 6237) is to be appended to it.

Alongside this legislative trajectory, the Brazilian government published in 2024 the Plano Brasileiro de Inteligência Artificial (PBIA 2024–2028), a national plan organised around five strategic axes. These include the development of national AI capacity, with sovereign large language models in Portuguese as a flagship initiative, AI for productivity and competitiveness, AI for the public sector, AI for inclusion and social welfare, and AI governance.[52] In functional terms, the PBIA occupies a position analogous to Australia's National AI Plan, although with a distinct emphasis on technological sovereignty, linguistic representativeness, and inclusive social policy. The Brazilian and Australian approaches thus combine, in each country, a legislative track and a policy-and-investment track, even as the relative weight and the substantive direction of each track differ markedly.

Few comparisons in contemporary AI regulation are as instructive as this one, precisely because the two trajectories begin from almost identical premises. Each country adopted a risk-based approach as its conceptual frame, began with ethical principles before contemplating binding rules, and organised broad public consultations as a constitutive part of its regulatory design. Each faced, in the late stages, sustained pressure from technology-sector actors arguing against AI-specific rules. The outcomes, however, could not differ more sharply. Brazil has continued to ad-

[52] BRASIL, Ministério da Ciência, Tecnologia e Inovações. *Plano Brasileiro de Inteligência Artificial 2024–2028 (PBIA): IA para o Bem de Todos*. Brasília: MCTI, 2024. Available at: https://www.gov.br/mcti/pt-br/acompanhe-o-mcti/transformacaodigital/plano-brasileiro-de-inteligencia-artificial. Accessed: 18 May 2026.

vance toward a binding statute coordinated by a single regulatory authority; Australia, after years of careful preparation, retreated to a technology-neutral model resting on pre-existing statutes and voluntary guidance. The shared starting points make the divergence itself the analytical object: it asks what conditions, beyond regulatory design, determine whether ambition is sustained or reversed.

From this divergence, Brazil may draw constructive lessons. Australia produced an unusually sophisticated body of operational material, including the twenty-six implementation practices, the National Framework for the Assurance of AI in Government, and the Guidance for AI Adoption, that translates abstract principles into actionable governance even where the underlying legislative project ultimately faltered. Such instruments offer practical templates that Brazilian regulators may draw upon when developing implementing regulations under whatever statutory framework is ultimately adopted. The Australian experience also highlights the strategic role of voluntary instruments as preparatory infrastructure, capable of building regulatory literacy in the technology sector and gathering practical evidence on what binding obligations should look like, even when the transition from voluntary to mandatory rules does not occur as planned.

Brazil's institutional design also offers elements that may inform the Australian debate, and indeed the broader international conversation. The most significant of these is the design of the SIA itself. Australia's post-2025 model does not entirely lack a dedicated AI institution: the AISI is expected to monitor, test, and share information on emerging AI capabilities, risks, and harms, and to advise government on areas of regulatory reform and capability gaps. Yet the AISI is an advisory body, without enforcement or coordinating powers over sectoral regulators, and in this it differs from the institutional role contemplated for the ANPD within the Brazilian framework. Where Australia continues to rely primarily on a constellation of pre-existing specialist regulators and various sectoral bodies (OAIC, ACCC, eSafety), the Brazilian design positions the ANPD as the coordinating hub of a network of sectoral authorities, with the CBIA providing strategic policy direction. Its significance lies not in centralising all AI oversight in a single agency, but in attempting to reconcile cross-sectoral coherence with the technical expertise of existing sectoral regulators.

This architecture addresses precisely the coordination problem that critics of the Australian model have identified, namely the risk that technology-neutral statutes

applied by dispersed regulators will produce inconsistent treatment of similar AI-related harms across sectors. The AISI may partially mitigate this problem by generating technical expertise and advising government on emerging risks, but its consultative character limits its capacity to ensure regulatory consistency across the system. The Brazilian choice to build coordination around an existing data protection authority follows an institutional logic with international resonance: the French CNIL played an analogous role in early European debates on AI oversight, and the European Union's emerging AI Office reflects similar coordinating ambitions. The Brazilian approach has the additional advantage of grounding AI governance in an authority that already operates with technical expertise on automated decision-making, algorithmic transparency, and the rights of affected persons.

5. CONCLUDING REMARKS

The regulation of artificial intelligence has emerged, over the past decade, as one of the central regulatory challenges facing contemporary democracies. The debate cuts across questions of innovation policy, fundamental rights, institutional design, international competitiveness, and the very capacity of states to govern transformative technologies whose pace of development outstrips the rhythms of legislative deliberation. Within this landscape, the comparison between Australia and Brazil offers an instructive case study, not because either country provides a model for the other to follow, but because the divergence of their trajectories illuminates the contingencies and pressures that shape AI regulation in constitutional democracies.

The Australian case illustrates both the possibilities and the fragility of regulatory ambition in this field. A carefully prepared project, supported by extensive consultation, expert input, and detailed technical groundwork, did not translate into a binding regime, yielding instead to a technology-neutral arrangement complemented by voluntary guidance and an advisory AI Safety Institute. The Brazilian case, by contrast, illustrates the slow but sustained advance of a binding, rights-anchored framework, anchored in a digital-rights tradition and articulated around the ANPD as coordinating authority within the SIA. Whether that commitment will be carried through to enactment, and on what terms, still remains an open question.

Taken together, the two trajectories invite a broader reflection on the conditions

that allow regulatory ambition in the AI field to be carried through. The depth of institutional embedding of the regulatory project shapes its resilience against late-stage reversal. The clarity and constitutional weight of the rights and principles that the regulation seeks to protect determine the normative anchor against which competing pressures must be weighed. The integration of the regulatory framework with pre-existing institutional capacity, particularly in the form of an authority equipped to oversee automated decision-making and the protection of fundamental rights, lends technical credibility and political stability to the regime. And the legibility of the regulation in international terms, in dialogue with comparable frameworks elsewhere, contributes to the legitimacy and effectiveness of its implementation.

Two paths, then, but no single destination. Australia's trajectory of preparation and retreat and Brazil's trajectory of slow but sustained advance illustrate that the global recalibration of AI regulation is not a uniform tide carrying all jurisdictions in the same direction. It is, rather, a landscape of choices made under broadly similar economic and institutional pressures, yet producing substantially different results. The comparative perspective offers no recipe. It offers, instead, a clearer view of what is at stake when democracies decide how to govern a technology that is reshaping the conditions of public and private life.

REFERENCES

ARBEL, Yonathan; TOKSON, Matthew; LIN, Albert. Systemic Regulation of Artificial Intelligence. Arizona State Law Journal, Phoenix: Arizona State University, v. 56, n. 2, p. 545–620, Summer 2024.

AUSTRALIA, Department of Finance. National Framework for the Assurance of Artificial Intelligence in Government. June 2024. Available at: https://www.finance.gov.au/government/public-data/data-and-digital-ministers-meeting/national-framework-assurance-artificial-intelligence-government. Accessed: 12 May 2026.

AUSTRALIA, Department of Industry, Science and Resources. Australia's AI Ethics Principles. Available at: https://www.industry.gov.au/publications/australias-artificial-intelligence-ethics-principles/australias-ai-ethics-principles. Accessed: 12 May 2026.

AUSTRALIA, Department of Industry, Science and Resources. Australia to Establish New Institute to Strengthen AI Safety. Press release, 25 November 2025. Available at: https://www.industry.gov.au/news/australia-establish-new-institute-strengthen-ai-safety. Accessed: 12 May 2026.

AUSTRALIA, Department of Industry, Science and Resources. Developing the AI Ethics Principles.

Available at: https://www.industry.gov.au/publications/australias-artificial-intelligence-ethics-principles/developing-ai-ethics-principles. Accessed: 12 May 2026.

AUSTRALIA, Department of Industry, Science and Resources. National AI Plan 2025. Canberra: DISR, 2 December 2025. Available at: https://www.industry.gov.au/sites/default/files/2025-12/national-ai-plan.pdf. Accessed: 12 May 2026.

AUSTRALIA, Department of Industry, Science and Resources. Safe and Responsible AI in Australia: Discussion Paper. June 2023. Available at: https://storage.googleapis.com/converlens-au-industry/industry/p/prj2452c8e24d7a400c72429/public_assets/Safe-and-responsible-AI-in-Australia-discussion-paper.pdf. Accessed: 12 May 2026.

AUSTRALIA, Department of Industry, Science and Resources. Safe and Responsible AI in Australia Consultation: Australian Government's Interim Response. January 2024. Available at: https://storage.googleapis.com/converlens-au-industry/industry/p/prj2452c8e24d7a400c72429/public_assets/safe-and-responsible-ai-in-australia-governments-interim-response.pdf. Accessed: 12 May 2026.

AUSTRALIA, Department of Industry, Science and Resources. Safe and Responsible AI in Australia: Proposals Paper for Introducing Mandatory Guardrails for AI in High-Risk Settings. September 2024. Available at: https://storage.googleapis.com/converlens-au-industry/industry/p/prj2f6f02ebfe6a8190c7bdc/page/proposals_paper_for_introducing_mandatory_guardrails_for_ai_in_high_risk_settings.pdf. Accessed: 12 May 2026.

AUSTRALIA, Department of Industry, Science and Resources. Testing the AI Ethics Principles. Available at: https://www.industry.gov.au/publications/australias-artificial-intelligence-ethics-framework/testing-ai-ethics-principles. Accessed: 12 May 2026.

AUSTRALIA, Department of Industry, Science and Resources. Voluntary AI Safety Standard. September 2024. Available at: https://www.industry.gov.au/publications/voluntary-ai-safety-standard. Accessed: 12 May 2026.

AUSTRALIA, Department of Industry, Science and Resources; National AI Centre. Guidance for AI Adoption. October 2025. Available at: https://www.industry.gov.au/publications/guidance-for-ai-adoption. Accessed: 12 May 2026.

AUSTRALIA, Digital Transformation Agency. Policy for the Responsible Use of AI in Government. Available at: https://www.digital.gov.au/policy/ai/policy. Accessed: 12 May 2026.

AUSTRALIA, Parliament. Senate. Select Committee on Adopting Artificial Intelligence (AI). Adopting Artificial Intelligence (AI): Report. Canberra: Parliament of Australia, November 2024. Available at: https://www.aph.gov.au/Parliamentary_Business/Committees/Senate/Adopting_Artificial_Intelligence_AI/AdoptingAI/Report. Accessed: 12 May 2026.

AUSTRALIA. Productivity Commission. Harnessing Data and Digital Technology: Interim Report. Canberra: Productivity Commission, August 2025. Available at: https://www.pc.gov.au/inquiries/current/data-digital. Accessed: 12 May 2026.

AYRES, Ian; BRAITHWAITE, John. Responsive Regulation: Transcending the Deregulation Debate. Oxford: Oxford University Press, 1992.

BALDWIN, Robert; CAVE, Martin; LODGE, Martin. Understanding Regulation: Theory, Strategy, and Practice. 2nd ed. Oxford: Oxford University Press, 2012.

BRASIL. Lei nº 12.965, de 23 de abril de 2014. Estabelece princípios, garantias, direitos e deveres para o uso da Internet no Brasil (Marco Civil da Internet). Available at: https://www.planalto.gov.br/ccivil_03/_ato2011-2014/2014/lei/l12965.htm. Accessed: 12 May

2026.

BRASIL. Lei nº 13.709, de 14 de agosto de 2018. Lei Geral de Proteção de Dados Pessoais (LGPD). Available at: https://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/l13709.htm. Accessed: 12 May 2026.

BRASIL, Câmara dos Deputados. Comissão Especial destinada a proferir parecer ao Projeto de Lei nº 2338, de 2023, sobre Inteligência Artificial. Available at: https://www.camara.leg.br/proposicoesWeb/fichadetramitacao?idProposicao=2487262. Accessed: 18 May 2026.

BRASIL, Câmara dos Deputados. Projeto de Lei nº 6237, de 2025. Institui o Sistema Nacional para Desenvolvimento, Regulação e Governança de Inteligência Artificial — SIA. Brasília, 8 December 2025. Available at: https://www.camara.leg.br/proposicoesWeb/fichadetramitacao?idProposicao=2594000. Accessed: 18 May 2026.

BRASIL, Ministério da Ciência, Tecnologia e Inovações. Estratégia Brasileira de Inteligência Artificial (EBIA). Brasília: MCTI, 2021. Available at: https://www.gov.br/mcti/pt-br/acompanhe-o-mcti/transformacaodigital/inteligencia-artificial. Accessed: 12 May 2026.

BRASIL, Ministério da Ciência, Tecnologia e Inovações. Plano Brasileiro de Inteligência Artificial 2024–2028 (PBIA): IA para o Bem de Todos. Brasília: MCTI, 2024. Available at: https://www.gov.br/mcti/pt-br/acompanhe-o-mcti/transformacaodigital/plano-brasileiro-de-inteligencia-artificial. Accessed: 18 May 2026.

BRASIL, Senado Federal. Comissão de Juristas Responsável por Subsidiar Elaboração de Substitutivo sobre Inteligência Artificial no Brasil: Relatório Final. Brasília: Senado Federal, December 2022.

BRASIL, Senado Federal. Projeto de Lei nº 2338, de 2023. Dispõe sobre o uso da inteligência artificial. Available at: https://www25.senado.leg.br/web/atividade/materias/-/materia/157233. Accessed: 12 May 2026.

CAJUEIRO, Daniel O.; CELESTINO, Victor R. R. A Comprehensive Review of Artificial Intelligence Regulation: Weighing Ethical Principles and Innovation. Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4950727. Accessed: 12 May 2026.

DATA PRIVACY BRASIL RESEARCH. Audiências Públicas da Comissão Especial de Inteligência Artificial da Câmara dos Deputados: Dinâmicas Participativas e Temas Emergentes em 2025. São Paulo, October 2025. Available at: https://www.dataprivacybr.org/relatorio-audiencias-publicas-

da-comissao-especial-de-inteligencia-artificial-da-camara-dos-deputados-dinamicas-participativas-e-temas-emergentes-em-2025/. Accessed: 12 May 2026.

EBERS, Martin. Truly Risk-Based Regulation of Artificial Intelligence: How to Implement the EU's AI Act. 2023. Available at: https://ssrn.com/abstract=4496789. Accessed: 12 May 2026.

EUROPEAN COMMISSION. Digital Omnibus on AI Regulation: Proposal for a Regulation amending Regulations (EU) 2024/1689 and (EU) 2018/1139 as regards the simplification of the implementation of harmonised rules on artificial intelligence. COM(2025) 836 final, 2025/0359 (COD). Brussels, 19 November 2025. Available at: https://digital-strategy.ec.europa.eu/en/library/digital-omnibus-ai-regulation-proposal. Accessed: 18 May 2026.

EUROPEAN COMMISSION. Ethics Guidelines for Trustworthy AI. Available at: https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai. Accessed: 12 May 2026.

GOLDENFEIN, Jake; PARKER, Christine; WEATHERALL, Kimberlee. Australia's National AI Plan Has Just Been Released: Who Exactly Will Benefit? The Conversation, 3 December 2025. Available at: https://theconversation.com/australias-national-ai-plan-has-just-been-released-who-exactly-will-benefit-270981. Accessed: 12 May 2026.

IIS PARTNERS. Australia's National AI Plan: Big Vision, Missing Guardrails. 15 December 2025. Available at: https://www.iispartners.com/insights/2025/12/15/australias-national-ai-plan-big-vision-missing-guardrails. Accessed: 12 May 2026.

OECD. AI Principles. Available at: https://www.oecd.org/en/topics/ai-principles.html. Accessed: 12 May 2026.

PARENTONI, Leonardo. What Should We Reasonably Expect from Artificial Intelligence? SSRN Electronic Journal, v. XII, n. 02, p. 179–219, July 2022.

REID, Alistair; O'CALLAGHAN, Simon; LU, Yaya. Implementing Australia's AI Ethics Principles: A Selection of Responsible AI Practices and Resources. Sydney: Gradient Institute; CSIRO, 2023. Available at: https://www.industry.gov.au/publications/implementing-australias-ai-ethics-principles-selection-responsible-ai-practices-and-resources. Accessed: 12 May 2026.

SCIMEX. Expert Reaction: National AI Plan Shifts Away from Mandatory Guardrails. 2 December 2025. Available at: https://www.scimex.org/newsfeed/expert-reaction-national-ai-plan-shifts-away-from-mandatory-guardrails. Accessed: 12 May 2026.

SELZNICK, Philip. Focusing Organisational Research on Regulation. In: NOLL, Roger (ed.). Regulatory Policy and the Social Sciences. Berkeley: University of California Press, 1985.

SMUHA, Nathalie A. From a "Race to AI" to a "Race to AI Regulation": Regulatory Competition for Artificial Intelligence. Journal of Law, Innovation and Technology, v. 13, n. 1, 2021.

UNITED KINGDOM, Department for Science, Innovation and Technology. A pro-innovation approach to AI regulation. White Paper, CP 815. London: DSIT, March 2023. Available at: https://www.gov.uk/government/publications/ai-regulation-a-pro-innovation-approach. Ac-

cessed: 18 May 2026.

UNITED KINGDOM, Department for Science, Innovation and Technology. A pro-innovation approach to AI regulation: government response to consultation. London: DSIT, February 2024. Available at: https://www.gov.uk/government/consultations/ai-regulation-a-pro-innovation-approach-policy-proposals/outcome/a-pro-innovation-approach-to-ai-regulation-government-response. Accessed: 18 May 2026.

UNITED KINGDOM. International Scientific Report on the Safety of Advanced AI: Interim Report. London: GOV.UK, 2024. Available at: https://www.gov.uk/government/publications/international-scientific-report-on-the-safety-of-advanced-ai. Accessed: 12 May 2026.

UNITED KINGDOM. The Bletchley Declaration by Countries Attending the AI Safety Summit, 1–2 November 2023. London: UK Government, 2023. Available at: https://www.gov.uk/government/publications/ai-safety-summit-2023-the-bletchley-declaration/the-bletchley-declaration-by-countries-attending-the-ai-safety-summit-1-2-november-2023. Accessed: 12 May 2026.

UNITED STATES. Executive Order 14179 of January 23, 2025, Removing Barriers to American Leadership in Artificial Intelligence. Federal Register, v. 90, n. 20, p. 8741, 31 January 2025. Available at: https://www.federalregister.gov/documents/2025/01/31/2025-02172/removing-barriers-to-american-leadership-in-artificial-intelligence. Accessed: 18 May 2026.

REGULATING ARTIFICIAL INTELLIGENCE IN CHINA: A TIERED, AGILE AND EMPOWERING GOVERNANCE MODEL

Jiyu Zhang

Professor of Digital Law, Executive Dean of the Future Rule of Law Institute,

Renmin University of China

Jingzhou Sun

Professor of Digital Law, Renmin University of China

Yufei Ma

Postdoctoral Fellow, Renmin University of China

DOI: https://doi.org/10.59224/dti8.ch13

Abstract:

China has not yet enacted a unified, comprehensive artificial intelligence legislation equivalent to the European Union's AI Act. This does not mean, however, that China lacks AI regulatory frameworks. On the contrary, through multiple parallel approaches including prior formulation of ethical norms, application of general civil and administrative laws, targeted special regulatory rules, refined industrial self-regulation, local pilot initiatives, and supporting technical standards, China has preliminarily built a tiered, agile and empowerment-oriented AI governance model. Instead of centering on a single codified statute, this model interconnects general laws such as the Civil Code, Cybersecurity Law, Data Security Law and Personal Information Protection Law with sector-specific rules governing algorithmic recommendations, deep synthesis technology, generative artificial intelligence, autonomous driving, smart healthcare, intelligent finance and smart justice. It further translates abstract legal norms into tangible governance capacity via technical standards,

ethical review mechanisms, industrial self-governance and public oversight.

This paper distills three interlinked defining features of China's AI governance system. First, governing norms and participating stakeholders are tiered and diversified. Second, regulatory instruments adopt a problem-centered, risk-classified and dynamically adjustable approach, embodying the logic of agile governance. Third, governance objectives extend beyond risk containment to enhancing capacity for corporate compliance, governmental supervision, public participation and international cooperation, thereby fostering positive synergy between technological development and security safeguards. Building on this framework, this paper further analyzes the bridging function of technical standards between overarching legal principles and enterprise compliance practice, and discusses outstanding challenges confronting the current system, including inconsistent coordination across regulatory norms, insufficient public participation, inadequate legal remedies for infringed rights, questionable legitimacy of technical standards, and limited cross-border interoperability of governance rules.

Keywords: AI Regulation; Empowerment Oriented Governance; Multi-Stakeholder Co-Governance.

Table of contents: 1. Introduction: Beyond the Quest for a Unified Artificial Intelligence Act. 2. Normative Foundations of China's Artificial Intelligence Governance. 3. The Tiered Architecture of China's Artificial Intelligence Regulation. 4. Technical Standards: The Bridge Linking Legal Principles and Compliance Practice. 5. Pluralistic, Agile and Empowerment Oriented Governance Mechanisms. 6. Existing Challenges and the Global Significance of the Chinese Model. Conclusion. Main References.

1. INTRODUCTION: BEYOND THE QUEST FOR A UNIFIED ARTIFICIAL INTELLIGENCE ACT

International comparative studies on AI regulation often take the European Union's AI Act as a benchmark, based on which scholars examine whether other jurisdictions have enacted unified, comprehensive AI legislation with horizontal applicability. While this comparative lens delivers straightforward observations, it

may obscure the actual governance architectures across distinct legal regimes. To judge whether a country has established an effective AI regulatory system, one should not merely check if it possesses a statutory law entitled an "AI Act"; instead, it is necessary to examine how general laws, administrative regulations, ethical codes, technical standards, industrial governance, local pilot schemes, and law enforcement and judicial mechanisms function in concert.

China has not yet enacted a single comprehensive statute covering the entire lifecycle of AI, including its research and development, provision, deployment and utilization. Nevertheless, multi-layered norms have been formulated to continuously address issues arising from AI in the spheres of rights protection, ethics, security, market competition and industrial development. At the level of general laws, activities related to artificial intelligence are subject to the Civil Code, Cybersecurity Law, Data Security Law, Personal Information Protection Law, Copyright Law, Patent Law, Anti-Monopoly Law, Anti-Unfair Competition Law, and Consumer Rights and Interests Protection Law. In terms of special regulatory rules, China has successively rolled out provisions governing algorithmic recommendations, deep synthesis technology, generative artificial intelligence, and labelling for generative synthetic content. Within specific industries, scenario-specific norms and standards have been developed for autonomous driving, smart healthcare, intelligent finance, smart justice and other sectors. At the local level, regions including Shanghai and Shenzhen have explored coordinated pathways for AI governance and industrial growth via industrial promotion ordinances, pilot policies and regulatory sandboxes.

It merits clarification that this tiered regulatory framework is not merely a post-hoc descriptive label; it has clear policy origins. The State Council's New Generation Artificial Intelligence Development Plan proposed to "establish a framework of laws, regulations, ethics and morality to safeguard the sound development of artificial intelligence", alongside a three-phase strategic roadmap. The Opinions on Further Implementing the "AI+" Initiative, issued in August 2025, further mandates the "strengthening of policy, legal and regulatory safeguards". Against this backdrop, China's AI governance is marked by a distinctive approach that prioritizes ethical norms before advancing cross-domain legislation: science and technology ethics committees, industrial standards and policy documents first explore shared values and operational specifications, and only mature, stable, widely recognized practices

are subsequently elevated into ministerial rules or even statutory provisions. This gradual, iterative approach embodies the core logic of agile governance and holds the key to understanding why China's model adopts a multi-tiered structure rather than a single codified act.

Accordingly, China's AI regulatory regime is better conceptualized as a tiered governance architecture, rather than a system lacking unified comprehensive legislation. The core strength of this architecture lies not only in its abundant normative instruments, but also in the specialized division of labor between norms at different tiers and diverse governance tools: statutes lay down fundamental boundaries concerning rights protection, public security, liability allocation and regulatory competences; administrative and industrial rules target concrete risks and application scenarios; ethical codes set out guiding values and institutional governance requirements; and technical standards translate abstract legal mandates into test methodologies, performance indicators and operational workflows.

This paper further argues that China's governance model bears two additional characteristics: agility and empowerment. The term "agile" means that regulators do not wait for full technological maturity to introduce definitive arrangements via unified comprehensive legislation. Instead, they address emerging issues in a timely manner through narrow-scope targeted rules, interim measures, risk-based tiered classification management, iterative updating of technical standards, local pilot programs, and regulatory dialogue. Some scholars summarize the core of agile governance into four dimensions: "*rapid response and early intervention in the timing of governance; effective integration of flexible general principles and concrete categorized rules in normative design; interactive cooperation among all stakeholders in governance relations; and facilitative governance featuring fast procedural progress and mild intervention intensity as its regulatory approach* "[1]" *"Empowerment", by contrast, indicates that AI governance is not limited to prohibitions, restrictions and liability sanctions. It also seeks to boost the capacity of enterprises, governments and social entities to develop and deploy AI safely and credibly by means of public infrastructure,*

[1] ZHANG, Jiyu. Establishing a Multi-Stakeholder Co-Governance System for Algorithms. *Legal Science (Journal of Northwest University of Political Science and Law)*. v. 40, n. 01, p. 115-123. 2022.

open public data, intellectual property protection, technical standards, compliance support services and international collaboration. Centered on empowerment, scientific and effective governance enables a high level of AI industrial development while sustaining long-term secure and trustworthy operation of AI systems "[2]

This model differs fundamentally from two alternative paradigms: governance that relies solely on corporate self-regulation, and stringent ex ante control which treats artificial intelligence merely as a source of risks. Its core proposition holds that AI security is not a static condition, but a capacity that requires continuous cultivation amid technological, institutional and organizational advancement. China's Cybersecurity Law and Data Security Law respectively define cybersecurity as "a state of stable and reliable operation" and data security as "the capacity to maintain a sustained secure state". This statutory formulation combining both "state" and "capacity" inherently reveals the underlying logic that development and security are not inherently contradictory; instead, a dynamic balance between the two can be achieved through the rule of law, technical standards and multi-stakeholder collaboration.

2. NORMATIVE FOUNDATIONS OF CHINA'S ARTIFICIAL INTELLIGENCE GOVERNANCE

(I) Human-Centeredness and Beneficial AI

"*Human-centeredness serves as the starting value of China's artificial intelligence governance, and also stands as one of the rare consensus concepts shared by the international community on AI governance.* "[3]Almost all major countries and international organizations at the forefront of AI development and governance have adopted human-centeredness as their core principle. Examples include the Asilomar AI Principles issued in 2017; the Recommendation on the Ethics of Artificial Intelligence adopted at the 41st General Conference of UNESCO in 2021; the Bletchley

[2] ZHANG, Jiyu. The Establishment of Concepts and Institutional Construction of Empowerment-Oriented AI Governance. *China Legal Science*. n. 05, p. 61-81. 2024.

[3] ZHANG, Jiyu. The Establishment of Concepts and Institutional Construction of Empowerment-Oriented AI Governance. *China Legal Science*. n. 05, p. 61-81. 2024.

Declaration signed by China, the United States and 26 other countries and regions in the United Kingdom in 2023; and the UN General Assembly resolution on strengthening international cooperation in AI capacity-building, co-sponsored by China and unanimously adopted at the 78th UN General Assembly in 2024. Article 1 of the Global AI Governance Initiative released by China on 18 October 2023 states clearly at the outset: "The development of artificial intelligence should adhere to the human-centered philosophy, aim to advance the common well-being of humanity, take safeguarding social security and respecting human rights and interests as prerequisites, and ensure that artificial intelligence always evolves in a direction conducive to the progress of human civilization."

The advancement of AI technologies shall ultimately serve human dignity, human rights and all-round human development, rather than reducing human beings to passive subjects subject to technological systems and automated decision-making. This principle is embedded in China's ethical documents, policy plans and specific legislation, and its legal implementation unfolds along three tiers. First, it centers on human rights to respect and protect fundamental human entitlements, manifested in safeguards for personal dignity, privacy, personal information, portrait rights, voice rights and related interests, a spirit fully embodied in statutes including the Civil Code and the Personal Information Protection Law. Second, it embodies fairness and justice by prioritizing algorithmic equity and digital accessible design, striving to narrow the digital divide and realize inclusive development. It mandates transparency, impartiality and equity in automated decision-making, and reserves the right to explanation, the right to reject fully automated outcomes, and avenues for human intervention where automated decisions exert substantial impacts on individuals. Third, it upholds human subjectivity and autonomy throughout AI operation via human-in-the-loop oversight and human-machine collaborative control mechanisms, guaranteeing meaningful human supervision and authority over AI systems. In high-stakes domains such as healthcare and judicial proceedings, AI is confined to an auxiliary role, and core decisions involving public authority or professional judgment must never be fully delegated to machines.

"*Beneficial AI represents the further extension and concretization of the human-centered philosophy within the artificial intelligence sector, as well as the embodiment of the principles of beneficial technology and beneficial digitalization for AI tech-*

nologies"[4]. Far from an abstract moral slogan, this principle requires individuals, enterprises, industrial practitioners and all other stakeholders engaged in AI research, development, provision and deployment to abide by public morality, core socialist values and the shared values of all humanity. It demands that intelligent technologies satisfy people's aspirations for a better life and continuously advance the common well-being of humankind. Meanwhile, it urges government authorities and social organizations to govern AI in a benevolent manner, guard against the malicious development and deployment of AI technologies, bridge the digital divide, and advance digital justice and social equity.

Realizing this philosophy requires institutional endeavors on two levels. First, guided by the human-centered and development-oriented core tenet, stakeholders shall continuously deepen their understanding of the value objectives governing AI application sectors, and embed the goal of "benevolence" into the entire R&D lifecycle. Second, concrete institutional mechanisms shall be leveraged to foster the innovative adoption of technical and managerial instruments that facilitate value alignment, safety and trustworthiness. Special emphasis shall be placed on rectifying the skewed research priorities of enterprises: without appropriate constraints and guidance, corporate R&D resources tend to flow disproportionately into areas promising tangible future profits, while insufficient investment is directed toward the protection of vulnerable groups, digital accessibility technologies and security safeguards. To address this imbalance, stronger guidance via ethical norms and legal institutions is indispensable.

Accordingly, AI governance must not merely evaluate systems in terms of accuracy, efficiency and robustness. It shall also examine whether such systems produce adverse outcomes including discrimination, manipulative behavior, privacy infringements, harm to labor rights, or the marginalization of vulnerable populations.

(II) Balancing Development and Security

The second normative foundation of China's artificial intelligence governance lies

[4] ZHANG, Jiyu. The Establishment of Concepts and Institutional Construction of Empowerment-Oriented AI Governance. *China Legal Science*. n. 05, p. 61-81. 2024.

in balancing development and security. Artificial intelligence is both a potential source of risks and a pivotal technology that fuels industrial upgrading, optimizes public services, and elevates national governance capacity. If governance solely targets risk elimination, it may stifle innovation by locking technological trajectories prematurely, raising market access costs, exacerbating uncertainties over liability, and narrowing room for regulatory trials. Conversely, unregulated market expansion of AI may exacerbate a host of hazards, including personal information leakage, algorithmic discrimination, synthetic disinformation, cybersecurity threats, market monopolies, labor manipulation, and public safety incidents. Therefore, development and security are not a simplistic binary opposition; instead, they constitute two intertwined imperatives that AI governance must address concurrently.

From a developmental perspective, artificial intelligence serves as a core driving force behind the new round of technological revolution and industrial transformation, marked by its fundamental and general-purpose nature. Its impacts extend far beyond internet platforms and digital content industries, permeating manufacturing, healthcare, transportation, finance, education, judicial administration, government services, scientific research, national defense and other sectors. AI innovation hinges on a multi-faceted ecosystem encompassing data, algorithms, computing power, talent pools, real-world application scenarios, a sound business environment, and intellectual property protection. Inadequate institutional support—such as ambiguous rules governing data circulation, underutilization of public data, unclear boundaries of intellectual property protection, and inconsistent regulatory expectations—can dampen enterprises' incentives to sustain investment in technological innovation and compliance solutions.

From a security perspective, AI risks feature public attributes, large-scale impacts and technical complexity. Such hazards are rarely private risks that individual users can fully identify, negotiate and mitigate on their own. Instead, they represent systemic risks capable of affecting massive groups of users, market order, public opinion ecology, public services and national security. For instance, deep synthetic content may infringe personality rights and erode social trust; generative artificial intelligence may generate false information and infringing materials; algorithmic recommendation systems may reshape public access to information and social cognition; autonomous driving and medical AI, meanwhile, bear direct bearing on human life

and health. For this reason, AI risks cannot be fully addressed merely through market self-regulation or individual consent; joint intervention via legislation, administrative supervision, technical standards and public oversight is indispensable.

The core of balancing development and security lies in framing security as an enduring capacity built upon technological progress, rather than a static state achieved by freezing technological advancement. High-quality development enables better risk prevention through advanced security technologies, evaluation tools, regulatory technology and enterprise compliance capabilities. In turn, robust security safeguards furnish a foundation of trust and stable expectations for technological deployment. Accordingly, AI governance must draw non-negotiable security bottom lines while reserving reasonable room for innovation, trial and error, and stable institutional expectations. The amendments to the Cybersecurity Law of the People's Republic of China, finalized in 2025 and effective in 2026 with newly added provisions governing artificial intelligence, precisely embody this balanced approach that attaches equal importance to basic theoretical research, core technologies, training datasets, computing infrastructure, ethical norms, risk monitoring and security supervision.

Accordingly, the principle of "equal emphasis on development and security" is not merely a policy slogan; instead, it constitutes the institutional logic running through tiered regulation, agile governance and empowerment-oriented governance. Tiered regulation prevents uniform regulatory stringency from being imposed across all scenarios. Agile governance avoids locking in future technological evolution through rigid rules at the early stage of technological development. Empowerment-oriented governance, by contrast, enhances the practical feasibility of secure and trustworthy AI development via capacity building.

(III) Inclusive and Prudent Regulation Coupled with Agile Governance

The principle of inclusive and prudent regulation advocates reserving reasonable room for trial and error for new technologies, products and business models, on the prerequisite that security bottom lines shall never be breached. It serves as a consistent, proven and advanced governance philosophy adopted by China in the science and technology sector, and has been embedded in national laws, administrative

regulations and ministerial rules. Article 35 of the Science and Technology Progress Law stipulates that "the State encourages the application of new technologies, and shall promote pilot trials of new technologies, products, services and models in accordance with the inclusive and prudent principle to create favorable conditions for their deployment". Article 3 of the Interim Measures for the Administration of Generative Artificial Intelligence Services explicitly imposes inclusive, prudent, tiered and classified supervision on generative AI services.

"*The core implication of this principle lies in adopting moderate tolerance toward emerging AI technologies, products and industrial forms. It allows relevant entities to rectify problems arising from research, development and deployment on their own, provided fundamental security thresholds are upheld. Government authorities conduct prudent follow-up monitoring and only intervene in a timely and proportionate manner, so as to avoid stifling nascent technological innovation through inappropriate regulation*"[5] . From the perspective of traditional regulatory theories, inclusive and prudent supervision reflects the law enforcement strategy of responsive regulation, the operational paradigm of cooperative regulation, and the evolutionary logic of experimentalist regulation.

"*Agile governance further elaborates the temporal structure and interactive modes of supervision* "[6]Its conceptual origin partially draws on agile development methodology from the software engineering sector of the 1990s. In 2018, the World Economic Forum reflected on policy-making amid the Fourth Industrial Revolution and formally put forward the concept of Agile Governance. It defines agile governance as adaptive, human-centred, inclusive and sustainable policy formulation, which entails sustained readiness to proactively or reactively embrace technological shifts and draw lessons from change .This concept has evolved from a promotional proposition into an institutional norm in China. In 2019, the National New Generation Artificial Intelligence Governance Specialist Committee issued Governance Principles for a

---

5 ZHANG, Jiyu. The Establishment of Concepts and Institutional Construction of Empowerment-Oriented AI Governance. *China Legal Science*. n. 05, p. 61-81. 2024.

6 ZHANG, Jiyu. The Establishment of Concepts and Institutional Construction of Empowerment-Oriented AI Governance. *China Legal Science*. n. 05, p. 61-81. 2024.

New Generation of Artificial Intelligence: Develop Responsible Artificial Intelligence, listing agile governance as one of eight core principles with detailed elaboration. In 2021, the same committee released Ethical Norms for New Generation Artificial Intelligence, which explicitly stipulates the promotion of agile governance within its administrative regulatory chapter. In 2022, the General Office of the Communist Party of China Central Committee and the General Office of the State Council jointly issued Guiding Opinions on Strengthening the Governance of Science and Technology Ethics, identifying agile governance as one of five core governance requirements. The document mandates enhanced early warning, tracking and assessment of sci-tech ethical risks, alongside dynamic, timely adjustments to governance modalities and ethical norms to swiftly and flexibly address ethical challenges stemming from technological innovation.

"*It is evident that agile governance has been repeatedly affirmed as a fundamental tenet of China's AI governance framework. Its core connotation consists of four dimensions: first, respecting the inherent developmental laws of AI while sustaining continuous monitoring and assessment, featuring rapid response and early intervention in governance timing; second, integrating flexible general principles with concrete categorical rules in normative design; third, prioritizing interactive collaboration across all stakeholders in governance relations; fourth, adopting facilitative governance characterised by streamlined procedures and mild regulatory intensity as its core intervention style.* "[7]

China's successive regulatory initiatives targeting algorithmic recommendations, deep synthesis technology and generative artificial intelligence in recent years serve as representative practical illustrations of agile governance. Instead of establishing complete, fully-fledged regulatory systems all at once, relevant rules have evolved incrementally alongside advances in technical functionality, rising social risks and accumulated governance experience. The scope of regulation has expanded sequentially from content recommendation systems to generative synthetic services, model safety requirements and synthetic content labelling obligations, forming an evolu-

[7] ZHANG, Jiyu. Establishing a Multi-Stakeholder Co-Governance System for Algorithms. *Legal Science (Journal of Northwest University of Political Science and Law)*. v. 40, n. 01, p. 115-123. 2022.

tionary regulatory chain that proceeds from principled declarations to concrete implementation, followed by periodic dynamic revisions. This trajectory will be elaborated further in Part III when discussing special sectoral regulation.

Institutionally, "a*gile governance is also manifested through the combination of flexible general clauses and specific categorical rules. On one hand, amid the rapid evolution of technology and incomplete legal frameworks, regulators rely to a certain extent on flexible normative anchors including social ethics and public morality to fill regulatory gaps. On the other hand, granular targeted provisions are formulated for well-defined concrete risks. For instance, the Provisions on the Management of Algorithmic Recommendations in Internet Information Services lays out detailed requirements respectively concerning algorithmic transparency, prohibitions on discriminatory user profiling, and bans on unreasonable differential treatment. This arrangement embodies the governance wisdom of coordinating general principles with specific rules .*"[8]

(IV) Capacity Building and Sustainable Development

"*Empowerment-oriented governance regards "capacity gaps" as the core challenge of artificial intelligence governance. At present, society generally suffers from insufficient capacity in multiple dimensions, including the cognition and assessment of AI risks, public oversight, internal corporate self-governance, and external governmental supervision. Three primary factors account for such inadequacies. First, AI systems feature black-box characteristics, partial autonomy, unpredictable outputs and unexplainable operational mechanisms, posing major obstacles to risk identification, evaluation, supervision and prevention. Second, AI has permeated nearly all social sectors. Comprehensive risk detection and assessment require joint participation not only from technical specialists, but also legal scholars, ethicists, sociologists and other experts to evaluate its holistic and scenario-specific impacts. Third, the widening digital divide creates stark disparities in the public's AI literacy and operational proficiency. This*

[8] ZHANG, Jiyu. Establishing a Multi-Stakeholder Co-Governance System for Algorithms. *Legal Science (Journal of Northwest University of Political Science and Law)*. v. 40, n. 01, p. 115-123. 2022.

hinders inclusive technological adoption and renders groups with limited AI understanding more vulnerable to associated hazards. "[9]Enterprises, particularly micro, small and medium-sized enterprises (MSMEs), often lack reserves of essential technical and managerial instruments, making it difficult to translate ethical codes and governance principles into tangible practices. Governments, meanwhile, are plagued by incomplete, asymmetric information concerning risk profiles and effective regulatory approaches, compounded by shortages of resources and dedicated tools. This undermines the operability, effectiveness and rationality of regulatory interventions.

Therefore, governance entails not only the allocation of obligations and liabilities, but also the development of foundational infrastructure covering training datasets, computing power, testing and evaluation mechanisms, certification systems, regulatory technology, public education and international collaboration. "*On this basis, the core essence of empowerment-oriented AI governance can be summarized into five dimensions of empowerment delivered through institutional governance* "[10]:First, empowering enterprises to pursue innovative development, especially boosting coordinated empowerment centered on data, computing, algorithm and institutional factors. Second, empowering enterprises to implement trustworthy regulation by popularizing risk management capabilities, enabling micro, small, medium and large enterprises to conduct effective internal self-governance. Third, empowering governments to deliver efficient supervision by developing and scaling up dedicated technologies for AI governance. Fourth, empowering the general public to equally share the benefits of AI advancement and enhancing public trust in artificial intelligence development. Fifth, empowering China to engage deeply in the construction of the global AI governance system. Sustainable secure and trustworthy development of artificial intelligence can only be achieved when innovation capacity, compliance capacity, governmental supervision capacity and public oversight capacity are upgraded in tandem.

Corresponding to capacity-building is the notion of sustainable development,

---

[9] ZHANG, Jiyu. The Establishment of Concepts and Institutional Construction of Empowerment-Oriented AI Governance. *China Legal Science*. n. 05, p. 61-81. 2024.

[10] ZHANG, Jiyu. The Establishment of Concepts and Institutional Construction of Empowerment-Oriented AI Governance. *China Legal Science*. n. 05, p. 61-81. 2024.

which likewise constitutes a natural extension of the core human-centered and development-oriented principles. "*Sustainable development stresses that technological advancement shall foster long-term health and balance across the economy, society and the environment. Its connotations encompass environmental friendliness, resource conservation, reduction of inequalities and the digital divide, advancement of education and employment, improvement in the quality and accessibility of public services, respect for and preservation of cultural diversity, promotion of open-access technology and knowledge-sharing, as well as support for global cooperation* ."[11] Sustainable development must also be pursued for AI technology itself, calling for rational approaches toward innovation and investment. The lessons drawn from the two historical "AI winters" serve as a reminder for regulators to prevent distorted resource allocation driven by either irrational exuberance or unwarranted pessimism. UNESCO's Recommendation on the Ethics of Artificial Intelligence advises governments and institutions to give full consideration to AI's impacts on the United Nations Sustainable Development Goals. China's Global AI Governance Initiative further proposes to actively leverage artificial intelligence for sustainable development so as to tackle global challenges such as climate change and biodiversity loss.

3. THE TIERED ARCHITECTURE OF CHINA'S ARTIFICIAL INTELLIGENCE REGULATION

(I) Ethical Norms and Science-and-Technology Ethics Review

China's early-stage AI governance was marked by a distinct feature of prioritizing ethical norms ahead of formal legislation. This approach rests on clear institutional bodies and a traceable chronological timeline. On 15 November 2017, under the leadership of the State Leading Group for Science and Technology Structural Reform and Innovation System Construction, the Office for the Promotion of the New-Generation Artificial Intelligence Development Plan was set up within the framework of the Inter-Ministerial Joint Conference on National Science and Technology Programme Management. It undertakes the organization and implementation of the

[11] ZHANG, Jiyu. The Establishment of Concepts and Institutional Construction of Empowerment-Oriented AI Governance. *China Legal Science*. n. 05, p. 61-81. 2024.

New-Generation Artificial Intelligence Development Plan and major national sci-tech projects. In 2019, this Office established the National New-Generation Artificial Intelligence Governance Specialist Committee, tasked with comprehensive research on AI-related policy frameworks, laws, regulations and ethical norms. On 24 July 2019, Xi Jinping presided over the 9th Meeting of the Central Committee for Deepening Overall Reform. The meeting reviewed and adopted the Plan for Establishing the National Science and Technology Ethics Committee. It clarified that science-and-technology ethics constitute value criteria that must be observed in all scientific and technological activities, and urged prompt improvement of institutional norms, refinement of governance mechanisms, and reinforcement of ethical supervision.

On this institutional foundation, the National New-Generation Artificial Intelligence Governance Specialist Committee successively issued Governance Principles for a New Generation of Artificial Intelligence: Develop Responsible Artificial Intelligence (2019) and Ethical Norms for New Generation Artificial Intelligence (25 September 2021). The latter sets out six fundamental ethical requirements: advancing human well-being, promoting fairness and justice, safeguarding privacy and security, ensuring controllability and trustworthiness, strengthening accountability, and enhancing ethical literacy. It further specifies eighteen concrete ethical requirements covering specific activities including AI administration, research-and-development, supply and deployment, integrating ethics and morality into the full lifecycle of artificial intelligence.In March 2022, the General Office of the CPC Central Committee and the General Office of the State Council issued the Guiding Opinions on Strengthening the Governance of Science and Technology Ethics, which advanced sci-tech ethics governance from declaratory norms toward institutionalisation. On 8 October 2023, ten authorities including the Ministry of Science and Technology, the Ministry of Education and the Ministry of Industry and Information Technology jointly promulgated the Trial Measures for Science and Technology Ethics Review. Formulated pursuant to the Science and Technology Progress Law and the above-mentioned Guiding Opinions, the document defines the main scope of science-and-technology ethics review, clarifies responsible entities, establishment criteria and operational mechanisms for ethics (review) committees, and stipulates con-

crete procedures including general review, simplified review, expert re-examination and emergency review. It expressly requires entities conducting scientific-technological activities in ethically-sensitive fields such as life sciences, medicine and artificial intelligence to set up science-and-technology ethics (review) committees.In 2025, ministries including the Ministry of Industry and Information Technology drafted the Trial Administrative and Service Measures for Science and Technology Ethics of Artificial Intelligence (Draft for Public Comment). It seeks to elaborate and implement sci-tech ethics governance requirements specifically for the AI sector and support enterprises in effectively strengthening their capacity to guard against science-and-technology ethical risks.

Ethical norms serve to establish value orientations, institutional procedures and professional judgement mechanisms in areas not yet covered by statutory law. Particularly for research activities touching upon human psychology, human life and health, capacities for social mobilisation, or highly-autonomous decision-making, ethics committees are able to identify risks at the early R&D stage and remedy the lag inherent in ex-post supervision. Nevertheless, ethical norms do not automatically carry the same coercive force as laws. Their effectiveness hinges on organisational accountability, regulatory alignment and accountability mechanisms. This issue will also be examined in the subsequent discussion concerning the soft-law nature of technical standards and the proposition of "hard-law and soft-law coordination".

(II) General-Purpose Laws as the Fundamental Protective Layer

China's artificial-intelligence-related regulation is not built from scratch. Existing laws already lay down fundamental constraints for AI-related activities, whose protective approaches can be roughly categorised into three dimensions: "right-based protection- conduct-based regulation- competition-order maintenance".From the perspective of right-based protection, the Personality Rights Book of the Civil Code addresses new forms of infringements arising in the intelligent era through provisions on portrait rights, voice rights and general personality interests. Article 1019 explicitly prohibits the infringement of others' portrait rights by means of distortion, defacement, or forgery via information-technology tools. Article 1023 sets forth

special protective provisions for voice-related interests in response to advances in voice-synthesis technology. Of greater forward-looking significance is Article 990 Paragraph 2: "In addition to the personality rights stipulated in this Book, natural persons enjoy other personality interests derived from personal liberty and personal dignity." This open-ended clause reserves room for legal interpretation for emerging personality interests that may emerge alongside AI development. The Personal Information Protection Law establishes mechanisms of transparency, fairness and impartiality for automated decision-making, together with individuals' rights to opt out and obtain explanations. It draws critical boundaries for AI systems built upon user profiling, content ranking, recommendation and credit evaluation. The Cybersecurity Law, the Criminal Law, the Civil Code and the Personal Information Protection Law jointly form a protective framework for personal-information-related rights and interests, empowering individuals to counter automated decision-making to a certain extent.

On the dimension of conduct-based regulation, the Data Security Law and the Cybersecurity Law govern AI R&D and services through provisions on classified and graded data protection, secure network operation, risk monitoring and national security concerns. The revised Cybersecurity Law of 2025, for the first time, introduces dedicated AI provisions at statutory level (Article 20), bringing algorithms, training datasets, computing-power infrastructure, ethical norms, security supervision and technological applications within one unified regulatory framework, as illustrated in Part II above. On the dimension of competition-order maintenance, the Anti-Monopoly Guidelines for the Platform Economy Sector, issued by the State Council Anti-Monopoly Committee on 7 February 2021, addresses horizontal monopoly agreements, vertical monopoly agreements and hub-and-spoke agreements concluded by leveraging data and algorithms. It also regulates abuses of dominant market position and algorithm-driven conducts that eliminate or restrict competition. The revision to the Anti-Monopoly Law, adopted by the Standing Committee of the National People's Congress on 24 June 2022, added Article 9 which stipulates that "Undertakings shall not engage in any monopolistic conduct prohibited by this Law by utilizing data and algorithms, technology, capital advantage, or platform rules, among others". For the first time, a fundamental competition statute explicitly

responds to algorithmic collusion and data-driven market dominance. Rules for personal-information protection also extend to concrete transaction scenarios. Article 18 of the E-Commerce Law of the People's Republic of China requires e-commerce operators, when delivering search results based on consumers' preferences and consumption habits, to simultaneously provide options not tailored to individual user characteristics. This represents one of the earliest direct legislative responses to recommendation and ranking algorithms. Furthermore, intellectual property law, copyright law and consumer-protection law respectively resolve issues such as right demarcation for training data and AI-generated content, platform dominance, as well as consumer rights to information and fair trade. Accordingly, AI regulation does not constitute a self-contained new branch of law; rather, it is a process in which multiple existing legal branches are reorganised and applied to emerging technologies.

(III) Targeted Special-Purpose Regulation for Specific Technologies and Risks

"Building upon general laws, China has gradually developed a set of special-purpose rules targeting particular technologies and risks. Their evolutionary process itself serves as a typical illustration of the agile-governance chain of "problem triggering-public deliberation-administrative rules-supporting standards".On 17 September 2021, nine central authorities including the Cyberspace Administration of China, the Central Publicity Department, the Ministry of Education, the Ministry of Science and Technology, the Ministry of Industry and Information Technology, the Ministry of Public Security, the Ministry of Culture and Tourism, the State Administration for Market Regulation and the National Radio and Television Administration jointly issued the Guiding Opinions on Strengthening the Comprehensive Governance of Algorithms for Internet Information Services. The document sets a roughly three-year timeframe to foster a comprehensive algorithm-security governance landscape featuring sound institutional mechanisms, a refined supervisory system and well-regulated algorithm ecosystems. It explicitly calls for multi-stakeholder algorithm-security governance mechanisms, and seeks to realise pluralistic co-governance characterised by governmental supervision, corporate accountability, industry self-regulation and public

oversight "[12]. On 31 December 2021, the Provisions on the Management of Algorithmic Recommendations in Internet Information Services was formally adopted and entered into force on 1 March 2022. Centering on challenges arising from algorithmic-recommendation services such as information cocoons, algorithmic black boxes, discriminatory user labels, unreasonable differential treatment, labour-rights protection and minor protection, it establishes mechanisms for algorithm filing, security assessment, user choice and complaint handling. Articles 10 and 18 lay down concrete provisions addressing algorithmic discrimination and algorithmic transparency respectively.

The Provisions on the Administration of Deep Synthesis in Internet Information Services further regulates generative-synthetic activities covering text, images, audio, video and virtual scenarios. It imposes obligations on service providers and technology supporters concerning security management, data processing and content labelling. Since the deployment of such technologies must not infringe personality rights, property rights, intellectual property rights or other legitimate rights, nor be deployed for illegal unfair competition, the portrait-right and voice-right protection clauses under Articles 1019 and 1023 of the Civil Code establish indispensable legal bottom lines for deep-synthesis technologies.

The Interim Measures for the Administration of Generative Artificial Intelligence Services, which entered into force on 15 August 2023, constitutes China's first special-purpose legislation dedicated to generative AI. It prescribes compliance with laws, administrative regulations, public morality and ethical norms for the provision and utilisation of generative-AI services, and clarifies the legal liabilities of service providers. Article 3 enshrines the principle of inclusive-prudent, tiered-and-classified supervision over generative-AI services, while Article 16 authorises competent state authorities to formulate tailored regulatory rules or guidelines adapted to the characteristics of respective industries and sectors. The Measures for Labelling of AI-Generated and Synthesised Content, promulgated in 2025 and effective September the same year, combines explicit and implicit labelling requirements. It mandates

[12] ZHANG, Jiyu. Establishing a Multi-Stakeholder Co-Governance System for Algorithms. *Legal Science (Journal of Northwest University of Political Science and Law)*. v. 40, n. 01, p. 115-123. 2022.

service providers to embed implicit markers within file metadata of AI-synthesised content in accordance with relevant provisions, and encourages implicit marking forms such as digital watermarks. Together with the supporting mandatory national standard Cybersecurity Technology-Labelling Methods for Artificial-Intelligence-Generated and Synthesised Content (GB 45438-2025, issued 28 February 2025), it forms an integrated institutional package. This sequential chain of rules illustrates that Chinese regulation generally defines the general direction of obligations via general laws and early-stage instruments first, before refining responsible parties, operational modalities and technical requirements in response to technological progress. Mature standards in turn drive further elaboration and improvement of supporting laws and regulations. This interactive relationship between laws and standards will receive dedicated analysis in Part IV.

(IV) Sector-Specific Regulation and Scenario-Based Governance

AI risks are heavily contingent upon application scenarios. The same algorithm yields markedly divergent risk profiles, scales of potential harm and tolerable margins of error when deployed for entertainment-oriented recommendation versus medical disease diagnosis. Accordingly, China's AI governance does not rely entirely on uniform horizontal rules. Instead, tailored scenario-specific requirements have been developed for sectors including healthcare, transportation, finance and judicial administration.

Within the healthcare sector, regulatory norms emphasise approval procedures, quality control, data management and personnel liabilities for AI-assisted diagnosis and treatment systems. Relevant regulatory attention dates back to an early stage. As early as 13 November 2009, the General Office of the former Ministry of Health issued the Trial Administrative Specifications for AI-Assisted Diagnostic Technology and the Trial Administrative Specifications for AI-Assisted Therapeutic Technology, establishing systems for follow-up visits, information retention and personnel competency assessment. In 2022, the National Health Commission and other authorities promulgated the Trial Detailed Rules for the Supervision of Online Medical Treatment, which further stipulates that AI software must not impersonate or replace physicians in delivering medical services, reaffirming the auxiliary role of artificial

intelligence.In the field of automated driving, the Ministry of Transport, the Ministry of Industry and Information Technology and the Ministry of Public Security jointly released the Trial Specifications for Road Testing of Intelligent and Connected Vehicles in April 2018. Multiple local governments subsequently rolled out detailed implementing rules for road testing and demonstration applications. On 24 March 2021, the Ministry of Public Security published the Draft Revised Proposal for the Road Traffic Safety Law, introducing new provisions governing the legal validity and safety of automated driving on public roads, road testing, driving-data recording, and liability determination for traffic accidents. At the technical-standard level, the State Administration for Market Regulation and the Standardization Administration of China issued the recommended national standard Classification of Automobile Driving Automation (GB/T 40429-2021) on 20 August 2021, which entered into force on 1 March 2022. It provides a classification benchmark for automated-driving administration, reflecting close mutual support between legislation and technical standards.

In the financial sector, robo-advisory and program trading are subject to financial regulatory rules. Article 23 of the Guiding Opinions on Regulating the Asset-Management Business of Financial Institutions, issued in 2018 by four authorities including the People's Bank of China, the former China Banking and Insurance Regulatory Commission, the China Securities Regulatory Commission and the State Administration of Foreign Exchange, sets out an extensive regulatory framework for investment-advisory services powered by artificial-intelligence technologies. The 2019 revision to the Securities Law added provisions governing program trading. In September 2023, under the guidance of the China Securities Regulatory Commission, the Shanghai, Shenzhen and Beijing Stock Exchanges formulated and issued the Circular concerning Matters Related to the Reporting of Stock Program Trading and the Circular on Strengthening the Administration of Program Trading. The former clarifies reporting subjects, modalities and content for program trading, with additional reporting requirements for high-frequency trading; the latter tightens procedural oversight over program-trading activities. In 2024, the China Securities Regulatory Commission further promulgated the Provisions on the Administration of Program Trading in the Securities Market (for Trial Implementation). Industry standards including Evaluation specification of artificial intelligence algorithm in

financial application (JR/T 0221-2021) and Guidance on information disclosure for financial applications based on artificial intelligence algorithms (JR/T 0287-2023), both approved and released by the People's Bank of China, further elaborate algorithm-evaluation methodologies and information-disclosure obligations. In the judicial domain, relevant opinions issued by the Supreme People's Court lay down core principles of safety and legality, fairness and impartiality, judicial-process support, as well as transparency and trustworthiness, embodying the fundamental position that artificial intelligence shall not replace judges in rendering judicial decisions in accordance with law.Recent judicial practice has accumulated rich adjudicative experience covering a range of issues: whether AI-generated content qualifies as works protected by copyright, liability attribution for copyright infringements caused by generative-AI systems, model training and the protection of voice-related rights and interests. Such precedents furnish a case-law foundation for future rule-refinement.

While scenario-based governance enhances regulatory pertinence, it may also give rise to overlaps and conflicts among norms issued by different administrative departments. Accordingly, establishing unified basic principles, risk-classification frameworks and liability interfaces constitutes a key challenge to be addressed by comprehensive future legislation, which will be further discussed in Part VI.

(V) Local-Level Experiments, Industrial Promotion and International Governance

At the level of local-level regulatory experiments, Shenzhen took the lead in enacting the Regulations on the Promotion of the Artificial Intelligence Industry in the Shenzhen Special Economic Zone in 2022. It stipulates that relevant municipal government authorities shall implement inclusive and prudent supervision for the AI industry in line with the principle of encouraging innovation. Corresponding regulatory rules and standards shall be formulated to accommodate the characteristics of new AI technologies, industries, business forms and models, alongside tiered-and-classified supervision. Meanwhile, the Regulations support and encourage scientific-technological research, development, market-oriented infrastructure construction and talent cultivation within the artificial-intelligence sector. Shanghai

has also promulgated the Shanghai Regulations on Promoting the Development of the Artificial Intelligence Industry. Local experiments make use of regional industrial foundations and regulatory resources to explore policy instruments adapted to emerging technologies, and accumulate practical experience for national-level legislation. Nevertheless, local normative documents must remain consistent with national laws. Regulators should guard against creating fresh market fragmentation and redundant compliance burdens.

At the international level, China's engagement in global artificial-intelligence governance has grown increasingly systematic in recent years. In 2023, at the 3rd Belt and Road Forum for International Cooperation, China put forward the Global AI Governance Initiative. It advocates the people-centered philosophy, the objective of "AI for good", the establishment of a risk-rating testing and assessment system, and the implementation of mechanisms for agile governance as well as tiered-and-classified administration. In the same year, China, together with 28 other countries and regions including the United States, the United Kingdom and the European Union, jointly signed the Bletchley Park Declaration at the inaugural AI Safety Summit. The document affirms that "artificial intelligence should be designed, developed, deployed and used in a safe manner for the benefit of all". On 1 July 2024, the 78th session of the United Nations General Assembly adopted by consensus the China-sponsored resolution on strengthening international cooperation on AI capacity-building (A/RES/78/311), co-sponsored by more than 140 countries. On 4 July of the same year, the Shanghai Declaration on Global AI Governance was issued at the 2024 World AI Conference and High-Level Meeting on Global AI Governance. In September 2024, at the High-Level Meeting on International Cooperation for AI Capacity-Building held at UN Headquarters in New York, China launched the AI for All Capacity-Building Initiative. Its priorities include advancing connectivity of AI and digital infrastructure, enabling thousands of industries through "AI Plus", improving AI literacy and talent development, and enhancing AI-related data security and diversity. On 26 July 2025, the Global AI Governance Action Plan was released at the 2025 World AI Conference and High-Level Meeting on Global AI Governance. It calls on all parties to jointly advance global AI development and governance based on goals and principles including AI for the people, respect for sovereignty, develop-

ment orientation, safety and controllability, fairness and inclusiveness, and open-door cooperation. Special emphasis is placed on bridging the intelligence divide and infrastructure development, positioning capacity-building as a core component of the global governance agenda. China's distinctive proposition within global governance holds that AI rules ought not merely focus on risk containment, but also secure access to technology, infrastructure and governance capacity for developing countries. This aligns closely in value orientation with the domestic concept of "empowerment-oriented governance" elaborated in this paper, jointly forming a holistic vision of the Chinese approach featuring mutual reinforcement between domestic and international dimensions.

4. TECHNICAL STANDARDS: THE BRIDGE LINKING LEGAL PRINCIPLES AND COMPLIANCE PRACTICE

(I) Dual Governance Demands for Artificial Intelligence and Functional Complementarity between Law and Standards

"*AI technical standards emphasise technical operability by specifying concrete technical methods, workflows, indicators and technical requirements, featuring timeliness, flexibility and forward-looking attributes. By contrast, laws and regulations set guiding principles, assessment procedures, supervision and accountability mechanisms. While endowed with authority and stability, they suffer from certain inherent laggards* ".[13] The two instruments are highly complementary in function. Technical standards concentrate on technical particulars, quantifiable requirements and operational norms for implementing ethical precepts. Their emphasis on practical operability fills the gap left by principled and abstract legal provisions. Rapid iteration of technical standards strikes a dynamic balance against the relative stability of legislation. Given that AI evolves far faster than legislative cycles, the flexible updating mechanism of technical standards effectively remedies the regulatory lag of statutory law. Furthermore, the high degree of internationalisation of technical standards

[13] ZHANG, Jiyu; MA, Yufei. Research on the Coordination Mechanism between Artificial Intelligence Technical Standard System and Legal and Regulatory System. *Admin. Reform*. n. 02, p. 119-128. 2026.

matches the demand for cross-border legal coordination. Harmonised international standards not only cut compliance costs for multinational enterprises, but also build technical consensus to underpin cross-border supervisory cooperation on artificial intelligence.

"*At present, China's AI technical-standard system has taken initial shape. At the international level, the Artificial Intelligence Subcommittee of the Joint Technical Committee 1 of the International Organization for Standardization / International Electrotechnical Commission (ISO/IEC JTC 1/SC 42) has released multiple standards including Artificial intelligence — Concepts and terminology (ISO/IEC 22989), Artificial intelligence — Knowledge engineering reference architecture (ISO/IEC 5392), Artificial intelligence — Application guidelines (ISO/IEC 5339), and Artificial intelligence management-system standard (ISO/IEC 42001). The Institute of Electrical and Electronics Engineers (IEEE) has also rolled out the IEEE 7000-series standards covering algorithmic transparency and ethical design. The European Commission issued the Coordinated Plan on Artificial Intelligence and set up the AI Standardisation Coordination Group. The U.S. National Institute of Standards and Technology (NIST) published the AI Risk Management Framework 1.0, which provides flexible, voluntary guidance. Domestically in China, four ministries including the Ministry of Industry and Information Technology jointly released the Guidelines for Building a Comprehensive National Standardization System for the Artificial Intelligence Industry (2024 Edition) in 2024. It sets targets of formulating more than 50 new national and industrial standards and participating in the development of over 20 international standards by 2026. In the same year, the National Technical Committee on Cybersecurity Standardisation released version 1.0 of the AI Security Governance Framework. To date, China has promulgated the mandatory national standard Cybersecurity Technology — Labelling Methods for Artificial-Intelligence-Generated and Synthesised Content (GB 45438-2025), alongside recommended national standards such as Artificial intelligence large-scale models — Part 2: Evaluation indicators and methods (GB/T 45288.2-2025) and Cybersecurity Technology — Basic Security Requirements for Generative-AI Services (GB/T 45654-2025). A multi-layered standard system consisting of national, industrial, local and association standards has been established across fields including automated driving, finance, healthcare and personal-informa-*

tion protection. By comparison, China's AI legal-regulatory system is undergoing accelerated legislation and framework construction. Grounded in the Cybersecurity Law, Data Security Law and Personal Information Protection Law, and supplemented by special administrative regulations such as the Interim Measures for the Administration of Generative Artificial Intelligence Services, it forms a governance regime dominated by principled and procedural requirements. It focuses on high-risk domains, features relatively long legislative cycles, and carries binding legal force ."[14]

(II) Manifestations and Underlying Causes of Insufficient Coordination

"*Nevertheless, coordination between the technical-standard system and the legal-regulatory system remains inadequate, giving rise in practice to several pressing challenges for improvement. Lacking operable technical support, legal provisions are difficult to implement on the ground. Certain risk-assessment standards reveal inherent limitations as technology evolves; unbacked by state coercive power, they suffer from implementation uncertainty. Enterprises are confronted with dual pressures: vague statutory requirements on the one hand and overly fragmented technical-standard requirements on the other. Supervision fails to keep pace with technical iteration. Technology advances faster than legislative proceedings, making it hard to respond promptly to evolving new-risk scenarios* "[15]. Specifically, insufficient coordination manifests itself in four major respects. First, there is inadequate linkage between the technology-oriented nature of technical standards and legal rules. Technical standards pursue precision, operability and quantifiability, whereas law pursues fairness, justice and right-definition and remains abstract and principled. Laws and regulations frequently define only risk-prevention objectives without specifying how such objectives can be achieved technically. This places enterprises in a dilemma: compli-

[14] ZHANG, Jiyu; MA, Yufei. Research on the Coordination Mechanism between Artificial Intelligence Technical Standard System and Legal and Regulatory System. *Admin. Reform*. n. 02, p. 119-128. 2026.

[15] ZHANG, Jiyu; MA, Yufei. Research on the Coordination Mechanism between Artificial Intelligence Technical Standard System and Legal and Regulatory System. *Admin. Reform*. n. 02, p. 119-128. 2026.

ance with standards does not guarantee satisfaction of legal obligations, yet compliance with the law offers little actionable technical guidance. Second, technical standards drafted mainly by industry and technical experts may become disconnected from laws and regulations. The legal status of non-mandatory standards remains ambiguous, and no clear rules govern the evidential weight of their assessment outcomes in administrative enforcement and judicial proceedings. Third, technical standards and legislation follow markedly different update cycles. For instance, China issued the recommended national standard Classification of Automobile Driving Automation back in 2021, yet revisions to the coordinating Road Traffic Safety Law took considerably longer.Fourth, tensions persist between international harmonisation of technical standards and divergent national legal regimes. For example, substantive divergences between the EU General Data Protection Regulation and China's Personal Information Protection Law may impose costly compliance-redesign burdens on multinational enterprises operating across jurisdictions.

"These observable symptoms stem from underlying causes spanning four dimensions: institutional, technological, value-based and interest-related."[16]From the institutional dimension, technical standards and laws belong to distinct institutional regimes. Technical standards are generally formulated by industry associations, leading enterprises, research institutes, university representatives and other stakeholders under the principle of voluntariness. Their effect derives from market recognition and industry self-regulation. By contrast, laws are enacted by legislative bodies through rigorous legislative procedures and backed by state coercive power. Such inherent institutional divergences result in the absence of proper linkage mechanisms between the two. From the technological dimension, artificial-intelligence technologies cover multiple fields including machine learning, natural-language processing and computer vision. Their large-scale deployment further gives rise to complexity emergence, a phenomenon in which simple components generate entirely new macro-level orders or properties through non-linear interactions, exhibiting the feature that "the whole is greater than the sum of its parts". This renders both

[16] ZHANG, Jiyu; MA, Yufei. Research on the Coordination Mechanism between Artificial Intelligence Technical Standard System and Legal and Regulatory System. *Admin. Reform*. n. 02, p. 119-128. 2026.

standards and laws hard to keep pace with technological progress in a timely manner. From the value dimension, technical standards prioritise technical feasibility, safety and efficiency and emphasise technical neutrality. The value system of law, however, encompasses structural values such as order, fairness and individual liberty. Divergences therefore emerge when value-laden trade-offs must be addressed, for instance the trade-off between algorithmic transparency and operational efficiency. From the interest-related dimension, industry interests may receive priority consideration during standard-setting, whereas law-making centres on public interests, social equity and social order. Tensions are thus intrinsic to their respective institutional objective structures.

(III) Soft-Law Attributes of Technical Standards and the Transformation Path of "Legal Principles-Technical Standards-Compliance Tools"

Addressing the relationship between technical-standard systems and legal-regulatory systems should not be confined merely to the adaptive gaps between soft law and hard law; instead, attention ought to be directed toward their collaborative co-governance. Soft law refers to effective rules of conduct that are not enforced by state coercive power . It may take diverse forms including codes of conduct, technical standards, moral statements, professional guidelines and certification programmes. Technical standards inherently embody the core characteristics of soft law: non-mandatoriness, voluntariness and flexibility. They are mostly drafted by non-governmental organisations, industry associations or technical standard-setting bodies. Their formulation builds upon technical consensus and interest reconciliation rather than formal state legislative procedures. Enterprises adopt standards largely out of market pressure, competitive strategies or brand-reputation considerations. Although technical standards lack statutory enforceability, they generate substantive guidance and restraints on market actors in practice. To improve product quality, gain access to international markets or mitigate compliance risks, enterprises often comply with relevant standards voluntarily, thereby giving rise to de-facto mandatory effects. Meanwhile, state administrative organs increasingly cite and incorporate standards within regulatory practice as important supervisory bases, realising a governance model whereby soft-law instruments reinforce hard-law

implementation.

The ideal coordination pathway lies in laws establishing inviolable rights-based bottom lines and regulatory objectives, which are then elaborated into measurable and enforceable requirements through technical standards, and ultimately translated into internal corporate policies, product designs and audit procedures. This constitutes the transformation path of "legal principles-technical standards-compliance tools".Article 16 of the Interim Measures for the Administration of Generative Artificial Intelligence Services exemplifies this logic: competent state authorities shall improve scientific regulatory approaches compatible with innovation and development, and formulate corresponding tiered-and-classified regulatory rules or guidelines in light of the technical features of generative-AI technology and its service applications in relevant industries and sectors. Legal principles and regulatory thresholds are defined first; technical standards and application specifications are then continuously refined, before reasonable and practical compliance tools are cultivated in the final stage. The labelling regime for AI-generated and synthesised content fully illustrates this process. Early normative documents first set forth the fundamental obligation that AI-synthesised content shall be labelled. The Measures for Labelling of AI-Generated and Synthesised Content further distinguishes between explicit and implicit labelling. Subsequently, the mandatory national standard GB 45438-2025, together with association standards and practical guidelines, specify concrete implementation methods such as file metadata embedding and digital watermarking. Through this chain, the principled labelling obligation evolves into implementable and inspectable technical workflows.

Once standards have been validated through practical implementation and built broad-based industry consensus, their mature provisions may be incorporated into statutes, administrative regulations or regulatory guidance via the mechanism of translating proven standards into legal norms. According to customary-law theory, long-standing, repetitive practices enjoying general recognition may serve as sources of law. As distillations of industry best practices, standards gradually evolve into industrial customs upon practical testing and sector-wide acceptance, thereby satisfying the prerequisites for elevation into legal norms. To this end, a mechanism for assessing and recommending standard maturity may be established. A standing

advisory body composed of legal, technical, industrial and ethical experts shall conduct periodic maturity assessments of relevant technical standards. Drawing upon standards such as Artificial intelligence — Service capability maturity assessment (GB/T 45907-2025), evaluation and scoring shall be carried out across dimensions including service capability, storage capacity, computing power, structured-data processing and decision-making capability. Core requirements of sufficiently mature standards may subsequently be elevated into statutory norms. Meanwhile, it should be recognised that standards cannot independently decide major rights restrictions or public-policy values. Matters touching upon fundamental rights, administrative penalties, market access and substantial allocation of liabilities must still be authorised and stipulated by laws and regulations. Greater transparency ought to be guaranteed throughout standard-setting procedures. Participation shall be expanded to include consumers, labourers, vulnerable groups, legal scholars and ethics experts. Mechanisms for objection, evaluation and periodic revision shall be put in place to prevent standard-setting from being dominated by a small number of large enterprises and thereby generating improper market-entry barriers.

(IV) Dynamic Coordination Mechanisms: Regular Communication, Sandbox Supervision and International Cooperation

"The relationship between laws and standards ought not to be a one-off, one-sided reference linkage. Instead, a dynamically operating system should be constructed. Through organic connection among regular communication frameworks, regulatory sandbox instruments and international-cooperation mechanisms, an iterative-adaptation ecosystem for technical standards and laws and regulations can be fostered, so as to address emerging risks brought forth by continuous technological shifts ."[17]

First, information-sharing and feedback channels constitute the core linkage of coordination mechanisms, whose key lies in establishing a regular communication system. Communicating parties should comprise not only standard-setting bodies

[17] ZHANG, Jiyu; MA, Yufei. Research on the Coordination Mechanism between Artificial Intelligence Technical Standard System and Legal and Regulatory System. *Admin. Reform*. n. 02, p. 119-128. 2026.

and legislative authorities, but also AI-related enterprises, researchers, industry organisations, the general public and other stakeholders. Building upon multi-stakeholder participation, user feedback mechanisms shall be refined and timely effect-evaluation procedures put in place.

Second, regulatory sandboxes represent an innovative pathway to realise dynamic coordination mechanisms. First introduced by the United Kingdom's Financial Conduct Authority (FCA) in November 2015, the regulatory sandbox provides enterprises with a supervised, safe testing environment. Applying this concept to the artificial-intelligence domain delivers a controllable testing ground for technological innovation. On the basis of regular information sharing and communication, it facilitates timely evaluation and adjustment of governance instruments and enables experimental practice of AI-oriented social governance. The EU AI Act requires Member States to ensure competent authorities establish at least one national-level AI regulatory sandbox. China's State Administration for Market Regulation, the Ministry of Industry and Information Technology and three other authorities jointly issued the Notice on Piloting the Automotive Safety Regulatory Sandbox Regime. It highlights the sandbox's values including enhancing emergency-response capacity, preventing and defusing major risks, safeguarding consumers' legitimate rights and interests, encouraging corporate technological innovation and advocating best-practice safety design. This furnishes an instructive institutional template for sandbox supervision within AI-relevant sectors.

Lastly, international-cooperation mechanisms for building AI governance capacity ought to be further advanced. Against the backdrop of divergent national legal systems, international standards serve as a "common language" for cross-border circulation and industrial collaboration. Harmonised technical benchmarks and collaborative rules reduce enterprises' costs of international operations while underpinning industrial development and global cooperation. This resonates with China's multi-layered engagement in global AI governance elaborated in Part III-(V). Collectively they shape a tripartite collaborative-governance landscape built upon "technical standards-legal rules-international norms".

5. PLURALISTIC, AGILE AND EMPOWERMENT-ORIENTED GOVERNANCE MECHANISMS

(I) Government Supervision: From Command-and-Control to Tiered-Classification and Continuous Interaction

The government remains the core organiser of China's artificial-intelligence governance. It is responsible for setting statutory bottom lines, allocating regulatory competences, enforcing administrative law and providing public infrastructure. Nevertheless, confronted with technical complexity and information asymmetry, effective governance cannot be achieved by ex-ante licensing or ex-post sanctions alone. "*Existing regulatory models include command-and-control regulation, persuasive regulation, responsive regulation and meta-regulation, which can be compared along dimensions such as regulators, regulatees, types of mandates and types of consequences. Command-and-control regulation demands high rule-specificity, which is difficult to realise amid the rapid evolution of artificial intelligence. Persuasive regulation fosters cooperation rather than confrontation, yet its effectiveness is often substantially undermined by corporate profit-driven incentives. Both approaches risk falling into the Collingridge Dilemma: at the early, still-controllable stage of technological development, harmful consequences are insufficiently understood to justify regulatory intervention; once adverse impacts become manifest, control turns costly and sluggish.*"[18]

"To address this challenge, China's algorithm and artificial-intelligence governance has gradually shifted toward a more comprehensive and dynamic regulatory paradigm grounded in responsive regulation and meta-regulation. It may be further expanded by drawing upon the "enforcement pyramid" model proposed by Ayres and Braithwaite. Regulatory instruments are arranged in tiers of ascending intervention intensity. Regulators are advised to prioritise light-touch interventions and adjust measures according to individual circumstances of regulatees. Building on this frame-

[18] ZHANG, Jiyu. Establishing a Multi-Stakeholder Co-Governance System for Algorithms. *Legal Science (Journal of Northwest University of Political Science and Law)*. v. 40, n. 01, p. 115-123. 2022.

work, three pyramids may be constructed for AI governance, with dynamic cross-tier adjustments informed by practical real-world information . "[19] The first is a risk-management pyramid featuring scientific tiered-classification and dynamic adaptability. It requires AI risk-management measures to match the level and category of risks. Targeted and effective controls are imposed on high-risk systems, while administrative burdens are reduced for numerous low-risk applications. This model is reflected in the four-tier risk taxonomy of "unacceptable risk-high risk-limited risk-minimal risk" under the EU AI Act, as well as the legislative logic of Article 23 of China's Provisions on the Management of Algorithmic Recommendations in Internet Information Services, which prescribes tiered-classified administration for algorithm-recommendation service providers. The second is a gradually-applied "regulatory-sanction pyramid". It follows the graduated set of regulatory sanctions laid down in the Cybersecurity Law, Data Security Law and Personal Information Protection Law, ranging from orders for rectification and formal warnings, through confiscation of illegal gains, fines, orders for business suspension and revocation of business licences, up to criminal liability. Proportionate sanctions prioritising light-weight intervention shall be adopted in light of concrete risk conditions. The third is a value-alignment incentive pyramid built upon effective interaction. Institutional incentives including administrative rewards, joint credit incentives, preferential credit treatment and tax benefits encourage enterprises to pursue technological R&D aligned with social values. It compensates for the insufficient incentives embedded within traditional regulatory mechanisms for investment in value-aligned, secure and trustworthy technological development. Combined deployment of these three pyramids enables regulation to maintain necessary deterrence for high-risk scenarios while reserving sufficient room for a broad range of innovative activities.

(II) Corporate Self-Governance and Industry Collaboration

"As developers, providers and deployers of AI systems, enterprises possess access to models, datasets and operational information. Accordingly, they bear primary respon-

[19] ZHANG, Jiyu. The Establishment of Concepts and Institutional Construction of Empowerment-Oriented AI Governance. *China Legal Science*. n. 05, p. 61-81. 2024.

sibilities for source-based governance. They may continuously advance scientific and technological innovation, identify and remedy algorithmic vulnerabilities and defects, and formulate internal rules to integrate technological and institutional strengths so as to implement the most direct and effective self-regulation."[20] Effective corporate self-governance shall cover risk-management regimes, ethical review, data governance, model testing, incident reporting, user complaint handling, human-in-the-loop intervention, and internal liability allocation. In terms of concrete institutional arrangements, four measures stand out. First, laws and regulations shall define risk management and prevention obligations for enterprises. Combining flexible clauses with concrete provisions shall encourage firms to adopt sound self-regulatory norms, complete management workflows and secure technical measures. For instance, the Provisions on the Administration of Algorithmic Recommendation for Internet Information Services mandates enterprises to establish management systems covering user registration, review of information release, algorithm-mechanism review, security assessment and monitoring, emergency response to security incidents, data-security protection and personal-information protection. Second, enterprises shall be encouraged to set up relatively independent and centralised algorithm risk-prevention committees or working groups with members of diversified professional backgrounds, tasked with coordinating and reviewing compliance-related matters. Third, corporate-compliance construction shall be promoted, including exploring the application of the corporate compliance non-prosecution system within the artificial-intelligence domain. Introducing corporate-compliance incentive mechanisms into public-prosecution procedures plays a positive role in preventing corporate offences, strengthening internal self-regulation and fostering law-abiding corporate operations. Fourth, corporate technological innovation shall be incentivised. Federated learning serves as a representative technical approach: algorithms are trained across multiple distributed edge devices or servers without exchanging raw data samples, thus addressing data-isolation and data-privacy-protection challenges. It constitutes a practical example where algorithmic innovation

[20] ZHANG, Jiyu. Establishing a Multi-Stakeholder Co-Governance System for Algorithms. *Legal Science (Journal of Northwest University of Political Science and Law)*. v. 40, n. 01, p. 115-123. 2022.

underpins the protection of individual rights, deserving further encouragement via patent regimes and policy guidance.

Nevertheless, corporate self-governance must not be interpreted as enterprises having authority to unilaterally set public rules. Subject to market competition, R&D costs and profit-seeking pressures, enterprises may overlook long-term risks and the interests of vulnerable groups. Furthermore, many AI firms face objective constraints such as insufficient computing power and scarcity of high-quality training data, which further limit their capacity for effective self-governance. Governments ought to constrain corporate self-governance through legal liabilities, information disclosure, external audits and regulatory inspections. For example, the Personal Information Protection Law stipulates that super-large internet platforms shall periodically issue social-responsibility reports on personal-information protection for public oversight. Meanwhile, standards, public testing platforms, compliance guidelines and support programmes for small-and-medium-sized enterprises shall be deployed to reduce compliance costs. Industry associations and open-source communities can also enhance risk-prevention capacity across the whole sector by formulating codes of conduct, sharing security information and developing evaluation tools. China's 14th Five-Year Plan specifically emphasises support for innovative consortia including open-source communities for digital technologies. Article 9 of the Data Security Law explicitly includes research institutes among multiple stakeholders in pluralistic co-governance, reflecting recognition of the role played by scientific-technological forces in data and algorithm governance.

(III) Social Supervision, Individual Rights and Remedies

AI systems exert far-reaching impacts on consumers, labourers, content creators and recipients of public-service delivery. News media, consumer organisations, research institutes and technical communities are capable of identifying systemic problems hardly detectable by individual end-users and pressing enterprises and regulators to take responsive measures. Several typical cases from China's recent algorithm-governance practice illustrate this mechanism.

One such example concerns the optimisation of news-recommendation algo-

rithms. On 6 July 2017, People's Daily published the article Do Not Let Algorithms "Kidnap" News. It drew public and central-government-agency attention to risks such as information cocoons and echo-chamber effects amplified by algorithmic recommendation. In January 2018, Toutiao disclosed its algorithmic mechanisms and solicited public comments. By prioritising important news and enabling multi-dimensional information exploration, it gradually embedded journalistic-ethical values into algorithm design. *"This practice was subsequently institutionalised in the Provisions on the Administration of Algorithmic Recommendation for Internet Information Services. The regulation imposes obligations on enterprises to proactively display content consistent with mainstream value orientations on key interfaces such as homepages and first-screen views, and to establish sound mechanisms for human intervention and user-autonomous choice ."*[21]

Second is the case concerning labour-rights protection for food-delivery riders. In 2020, the magazine Portrait published the article Food Delivery Riders Trapped in the System. It drew high-level attention from multiple central authorities including the State Administration for Market Regulation and the All-China Federation of Trade Unions. Jointly-issued Guiding Opinions on Fulfilling the Responsibilities of Online Catering Platforms and Effectively Safeguarding the Rights and Interests of Food-Delivery Workers followed. Major online food-delivery platforms subsequently invited multi-stakeholder representatives to deliberate algorithmic rules, disclosed the algorithms for estimated arrival-time calculation, and built-in human-intervention mechanisms for abnormal scenarios. Articles 11 and 17 of the Provisions on the Administration of Algorithmic Recommendation for Internet Information Services further lay down targeted rules requiring platforms to improve algorithm design and fulfil obligations for labour-right protection.

Third, lessons may be drawn from overseas experience. ProPublica, a United-States-based non-profit investigative journalism organisation, analysed the widely-used COMPAS risk-assessment tool deployed within U.S. judicial administration and uncovered clear racial bias against Black defendants. This finding drove the

[21] ZHANG, Jiyu. Establishing a Multi-Stakeholder Co-Governance System for Algorithms. *Legal Science (Journal of Northwest University of Political Science and Law)*. v. 40, n. 01, p. 115-123. 2022.

development of algorithm-impact assessment and accountability mechanisms for public-sector AI applications across the United States and Canada. The case underscores the necessity of valuing independent social-oversight actors as China advances governmental AI deployment.

Moreover, after Suzhou launched the "Sucheng Civilisation Code", extensive media coverage and scholarly commentary prompted local-government reflection and more prudent administrative practice regarding the boundaries of algorithm-driven governance.

Individuals may participate in governance by exercising rights to be informed, to choose, to opt out, to demand explanations, to file complaints, to make reports and to initiate litigation. Existing legislation already confers certain rights concerning automated-decision-making and personal-information protection. Examples include the right to opt out of automated decision-making under the Personal Information Protection Law, the "notice-and-take-down" regime established by the Regulations on the Protection of the Right to Information Network Communication and the Civil Code, as well as Article 42 of the Provisions on the Administration of Algorithmic Recommendation for Internet Information Services, which obliges algorithm-recommendation service providers to set up accessible complaint-and-reporting portals and maintain user-appeal channels and procedures. Nevertheless, in practice individuals still confront obstacles such as algorithmic opacity, evidential difficulties, dispersed harms and excessively high remedy costs. As observed in scholarship, "*ordinary citizens generally lack the cognitive capacity, preventive capacity and bargaining power to cope with modern public risks posed by algorithms. They can hardly make rational choices with full understanding of relevant risks, which undermines the normative foundation for risk-bearing by individuals (Zhang Jiyu, 2022).*"[22]

Going forward, further clarification is needed regarding the scope of explanatory obligations, substantive standards for human-in-the-loop intervention, platforms' duties of evidence preservation, alongside mechanisms for public-interest litigation

[22] ZHANG, Jiyu. Establishing a Multi-Stakeholder Co-Governance System for Algorithms. *Legal Science (Journal of Northwest University of Political Science and Law)*. v. 40, n. 01, p. 115-123. 2022.

and collective redress. Consideration may also be given to providing necessary immunity for "black-box testing" carried out for compliance- and ethics-oriented research purposes within copyright-related statutes. This would facilitate algorithm oversight by technical communities and the general public.

(IV) Empowering Innovation and Compliance-Building Capacity

Empowerment-oriented governance is not synonymous with simple industrial support. Instead, it creates institutional conditions enabling innovators to conduct responsible R&D within a predictable framework. This is consistent with the five-dimensional notion of "empowerment" elaborated in Part II. Its main components include improving rules for law-based data circulation and the development and utilisation of public data. Relevant policy documents comprise the Opinions of the CPC Central Committee and the State Council on Establishing a Basic Data System to Better Tap the Role of Data as a Factor of Production, the Opinions of the General Office of the CPC Central Committee and the General Office of the State Council on Accelerating the Development and Utilisation of Public Data Resources, and the Implementation Specifications for Authorised Operation of Public Data Resources (Trial) issued by the National Development and Reform Commission and the National Data Bureau. These furnish institutional underpinnings for inclusive empowerment driven by data factors. Second, computing-power infrastructure and testing-and-evaluation facilities shall be strengthened. Third, intellectual-property and competition-law instruments shall protect innovation while preserving open markets. For instance, the China National Intellectual Property Administration reinforces intellectual-property protection by refining and clarifying examination rules. It has added a dedicated section within the Patent Examination Guidelines governing examination of invention-patent applications involving algorithm features or features of business rules and methods, and released examination guidance for AI-related invention-patent applications. This delivers effective incentives for enterprises engaged in R&D on AI value-alignment mechanisms and regulatory technologies. Fourth, compliance-technology and regulatory-technology solutions shall be fostered. Fifth, standards, guidelines, certification services and public services shall be made available for small-and-medium-sized enterprises, relieving them of R&D costs

incurred by blind trial-and-error exploration.

"Equally important is the empowerment of governmental bodies and the general public. Regulators require interdisciplinary talent, technical testing tools, and capacity to obtain risk-related information on an ongoing basis. These constitute the organisational foundation for the real implementation of agile governance. The public, in turn, needs AI literacy, transparent information disclosure and accessible remedy channels. Pluralistic co-governance will remain merely formal unless all types of stakeholders possess practical capacity to participate in governance. This represents the concrete application of the vision of "joint construction, joint governance and shared benefits" within algorithmic and artificial-intelligence governance. Joint construction serves as the foundation, joint governance as the key, and shared benefits as the ultimate goal. The concept of a social-governance community in which everyone bears responsibility, fulfils obligations and enjoys benefits is equally applicable to the governance of digital and intelligent societies."[23]

6. EXISTING CHALLENGES AND THE GLOBAL SIGNIFICANCE OF THE CHINESE MODEL

(I) Norm Fragmentation and Hierarchical Coordination

The merit of tiered governance lies in its capacity to respond swiftly to concrete issues, yet it comes at the cost of abundant rules promulgated by diverse authorities with varying levels of legal validity. Enterprises may simultaneously be subject to general statutes, ministerial regulations, local rules, national standards, industrial standards and platform-specific norms, and struggle to determine the priority among conflicting requirements. This challenge has been fully illustrated in the discussions on sector-specific regulation and technical standards in Parts III and IV.

Going forward, comprehensive legislation or unified guiding principles shall clarify the definition of artificial intelligence, risk classification, fundamental rights, stakeholder liabilities, regulatory coordination and mechanisms for standard incor-

[23] ZHANG, Jiyu. Establishing a Multi-Stakeholder Co-Governance System for Algorithms. *Legal Science (Journal of Northwest University of Political Science and Law)*. v. 40, n. 01, p. 115-123. 2022.

poration. Reviewing China's AI legislation from a global comparative perspective, some scholars observe that "although China's current "small-scale, rapid, targeted" sector-specific legislative approach enhances responsiveness, it also brings about fragmentation within the normative system, which urgently demands systematic integration through comprehensive legislation ."[24] The function of comprehensive legislation is not to simply supersede existing sector-specific rules, but to establish shared concepts and institutional interfaces so as to furnish a unified foundation for domain-specific governance. Otherwise, excessive pursuit of a single code may sacrifice the agility of the existing system. This explains why this paper consistently emphasises "tiered governance" rather than "centralised governance" as the core feature of the Chinese model.

(II) Further Improvement Needed for Public Participation and Right Remedies

While pluralistic participation has been emphasised in China's artificial-intelligence governance, institutional practice is still largely driven by administrative authorities, industrial actors and technical experts. There remains room to enhance the participation of ordinary users, labourers, consumers and vulnerable groups in rule-making, standard-setting and risk-assessment procedures. This is closely linked to the structural tendency of "prioritising industry interests" in technical-standard formulation discussed in Part IV.

Furthermore, information disclosure does not automatically translate into effective oversight. Disclosed content shall relate to concrete decisions and risks, and be presented in forms comprehensible and actionable for users. For high-impact automated decision-making, stronger procedural safeguards, independent evaluation and accountability mechanisms ought to be established. This echoes the discussion in Part V concerning excessive remedy-cost barriers facing individuals seeking redress.

---

[24] DING, Xiaodong. China's Artificial-Intelligence Legislation in Global Comparative Perspective. *Journal of Comparative Law Studies*. n. 04, p. 51-66. 2024.

(III) Legitimacy of Technical Standards and Risks of Regulatory Capture

"Although technical standards enhance the operability of rules, they may conceal value-laden choices behind technical jargon. If standard-setting is dominated by a small number of large-scale enterprises, it may generate entry barriers favouring incumbent market players, or dilute rights-protection requirements in the name of commercial feasibility. Due to inherent attributes of soft law, there has traditionally been scope for improving public trust. Issues such as the credibility of standard-setting bodies and lack of direct enforceability remain to be addressed."[25] Accordingly, procedural transparency, conflict-of-interest management, pluralistic participation and ex-post evaluation for standard-setting shall be strengthened. Distinction shall be drawn among the different legal effects produced by recommended standards, mandatory standards and standards incorporated by legal reference. These constitute the core issues to be resolved by institutional designs advanced in Part IV, including the assessment-and-recommendation mechanism for standard maturity and regular communication frameworks.

(IV) Insights from China's Experience for the Global South

The international significance of the Chinese model lies not in delivering a set of rules ready for direct transplantation, but in demonstrating that no single institutional path exists for artificial-intelligence governance. For many developing countries, directly copying highly complex and costly comprehensive compliance regimes — such as the EU AI Act framework of "risk-based classification and full-life-cycle regulation" — may exacerbate technological dependence and market concentration. Tiered, agile and empowerment-oriented governance suggests that each country may build its institutional system incrementally by starting with general-rights pro-

[25] ZHANG, Jiyu; MA, Yufei. Research on the Coordination Mechanism between Artificial Intelligence Technical Standard System and Legal and Regulatory System. *Admin. Reform*. n. 02, p. 119-128. 2026.

tection, scenario-specific rules, technical standards and regulatory experiments, in light of its industrial foundation, administrative capacity and social risks. Meanwhile, global AI governance ought not focus merely on risk spill-overs; it must also address the unequal distribution of computing power, data, talent, infrastructure and security tools. These are precisely the concerns targeted by China's AI Inclusive Capacity-Building Initiative and the Global Action Plan on AI Governance. Capacity-building should constitute an essential component of global rules. Nevertheless, the "right to development" cannot serve as justification for lowering standards of rights protection and public accountability. Truly sustainable empowerment-oriented governance shall integrate technological inclusiveness, individual rights, public interests and democratic participation as shared governance objectives.

CONCLUSION

China's artificial-intelligence regulation should neither be characterised by the absence of unified comprehensive legislation, nor reduced to a mere collection of scattered rules. Its emerging institutional logic builds a multi-layered governance structure composed of general-purpose statutes, special regulations, industrial norms, local pilots, ethical review and technical standards. It addresses technological uncertainty through risk-based tiered classification, dynamic adaptation and pluralistic participation.

A salient feature of this system is its conception of AI governance as a combination of risk control and capacity-building. The rule of law shall not only set baseline safeguards for security and individual rights; it shall also deliver stable expectations for innovation, foster corporate compliance capacity, equip regulators with technical tools, and grant the public access to participation and remedies. Technical standards serve as a bridge between legal principles and engineering practice, yet their legitimacy must rest upon statutory authorisation, open procedures and social oversight.

Going forward, China's AI governance shall continue to strike a balance between unity and flexibility. On the one hand, higher-level legislation shall strengthen foundational concepts, rights protection, risk categorisation and regulatory coordination. On the other hand, the agility of special rules, industrial standards and local pilots to respond to technological shifts ought to be preserved. The ultimate goal is not merely

risk mitigation or industrial expansion, but to cultivate institutional capacity that sustainably advances the safe, reliable, fair, controllable and inclusive development of artificial intelligence. As China has repeatedly emphasised in international-governance fora, efforts shall be made to unlock AI's full potential while guaranteeing its security, reliability, controllability and fairness, so as to build an inclusive, open, sustainable, equitable, secure and trustworthy digital and intelligent future for all.

REFERENCES

ZHANG, Jiyu. Establishing a Multi-Stakeholder Co-Governance System for Algorithms. *Legal Science (Journal of Northwest University of Political Science and Law)*. v. 40, n. 01, p. 115-123. 2022.

DING, Xiaodong. China's Artificial-Intelligence Legislation in Global Comparative Perspective. *Journal of Comparative Law Studies*. n. 04, p. 51-66. 2024.

ZHANG, Jiyu; MA, Yufei. Research on the Coordination Mechanism between Artificial Intelligence Technical Standard System and Legal and Regulatory System. *Admin. Reform*. n. 02, p. 119-128. 2026.

ARTIFICIAL INTELLIGENCE IN CHINA: THE REGULATORY MOVEMENT AND ITS RESPECTIVE STAGES OF DEVELOPMENT

Matheus Antes Schwede

Ph.D. candidate in Law at the Federal University of Minas Gerais (UFMG) with a CAPES/DS scholarship. Master of Laws with a concentration in Human Rights from the Regional University of the Northwest of the State of Rio Grande do Sul (UNIJUÍ).

Giovanni Simão Triginelli

Master of Laws from the Federal University of Minas Gerais (UFMG). Holds a Postgraduate Specialization in Health Law from Dom Helder. Bachelor of Laws from the Federal University of Minas Gerais (UFMG).

DOI: https://doi.org/10.59224/dti8.ch14

Abstract: China is one of the major international players when it comes to AI. Therefore, the study was guided by the following question: what are the elements that make up the regulatory framework for artificial intelligence in the People's Republic of China? The hypothesis posited is that there are two stages: an initial stage responsible for developing recommendations, plans, and ethical standards, and another (the current stage) involving the creation of more specific and detailed laws. The main results identified some components of the initial stage and the current stage. Some characteristics of this regulatory structure were also analyzed. The research, which was exploratory in nature, employed a methodological approach using bibliographic and documentary research techniques, with a qualitative approach.

Keywords: Artificial Intelligence. Regulation. China.

Table of contents: 1) Introduction; 2) Methodology; 3) The First Stage of Artificial Intelligence Regulation in China; 4) The Second Stage of Artificial Intelligence Regulation in China; 5) Considerations on the Chinese Regulatory Model; 6) Final Considerations; 7) References.

1 INTRODUCTION

When it comes to artificial intelligence and technological development, China is one of the most prominent countries in the discussion, whether regarding its significant progress in developing new technologies or in the field of regulating their use. Thus, in light of the regulatory movement that has emerged in various countries due to concerns about the effects of the rapid advancement of artificial intelligence, this study aims to analyze the regulatory movement regarding AI in China.

The proposed study is guided by the following research question: What are the elements that make up the regulatory movement regarding artificial intelligence in the People's Republic of China?

To address this research question, the initial hypothesis is that China, seeking to achieve a prominent position on the global artificial intelligence stage across all sectors – production, development, and regulation – has consequently implemented two stages of technology regulation. The first stage is marked by a period during which various documents were drafted, such as strategic plans, recommendations, and ethical principles that served as the basis for the development and use of artificial intelligence systems within China. Similarly, these documents also served as the foundation for the second regulatory phase. This second phase, understood as the current phase, built upon the recommendations of the initial phase, leading to the development of various laws to establish the manner of use, citizens' rights, and the development of artificial intelligence in the country, shifting its focus away from mere regulations and introducing stronger normative provisions.

2 METHODOLOGY

To conduct this exploratory research, the methodological approach consists of using bibliographic and documentary research techniques, carried out by accessing content relevant to the studies, using a qualitative approach. The HeinOnline database was used as the selection criterion for articles to develop the study. Thus, on June 15, 2025, at 1:40 p.m., the following terms were entered into the website's search: "China Artificial Intelligence Regulation," with the "All databases" option selected and the search filtered by relevance. Consequently, the first 30 articles were selected, and the article "Workplace Artificial Intelligence Regulation in China: Between Efficiency and Social Stability" by Liu, Zhang, and Sui[1] was accessed via the linked DOI. However, two articles were unavailable and were not accessed. Furthermore, those not in English, Spanish, or Portuguese were disregarded, along with those that did not contribute to addressing the research question.

Additionally, articles available on SSRN were selected. The search was conducted using Advanced Search with the terms "Artificial Intelligence Regulation in China" on June 29, 2025, at 2:49 PM. The first 11 results were selected, and only those available as open access were used. In the "Sort By" field, the following was selected: Downloads, Descending. The inclusion criteria were that the article could help address the proposed problem and that it be in Portuguese, English, or Spanish. Others were disregarded.

An external article[2] from the database was selected to address some aspects of the proposed first stage, as it was found in one of the texts used and contained a well-structured body of information. Additionally, the website China Law Translate

---

[1] See LIU, Mingwei; ZHANG, Hao; SUI, Yi. Workplace artificial intelligence regulation in China: Between efficiency and social stability. *ILR Review*, vol. 77, no. 5, pp. 813–824, 2024. DOI: https://doi.org/10.1177/00197939241278956b. Available at: https://journals.sagepub.com/doi/abs/10.1177/00197939241278956b. Accessed on: June 15, 2025.

[2] See TRIOLO, Paul; GOODRICH, Jimmy. From Riding a Wave to Full Steam Ahead: As China's Government Mobilizes for AI Leadership, Some Challenges Will Be Tougher Than Others. *New America DigiChina Blog*, 2018. Available at: https://www.newamerica.org/cybersecurity-initiative/digichina/blog/riding-wave-full-steam-ahead/. Accessed on: June 26, 2025.

(https://www.chinalawtranslate.com/en/) was used to analyze relevant legislation when necessary.

To collect data on ethical principles, the Worldwide AI Ethics Dashboard tool was used on June 22, 2025. This tool allows users to explore 200 publications from around the world, enabling the selection of continent, region, country of origin, types of institutions, specific institutions, year of publication, and other information, as shown in Figure 1.

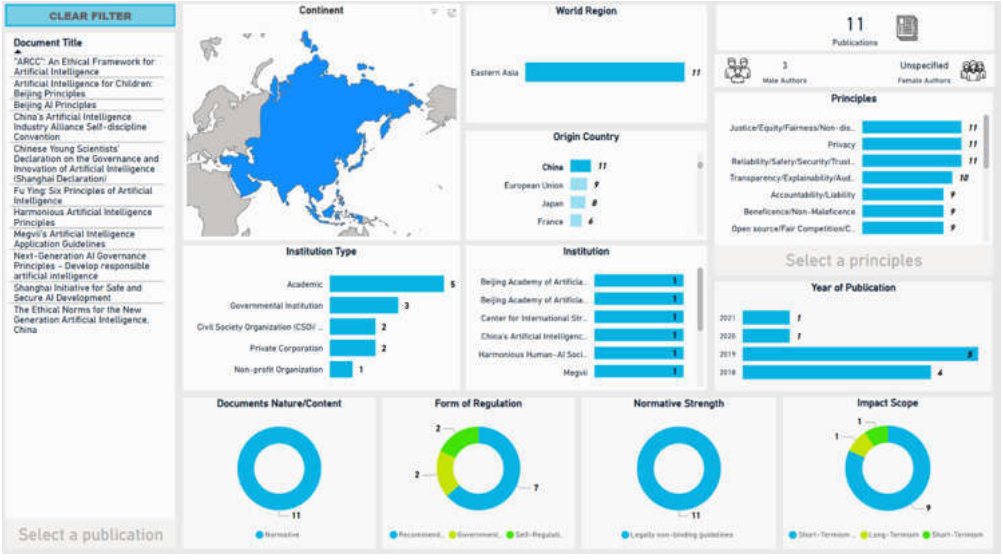
Figure 1 - Worldwide AI Ethics Dashboard



Source: CORRÊA, Nicholas Kluge et al. *Worldwide AI Ethics Dashboard*. 2023. Available at: https://nkluge-correa.github.io/worldwide_AI-ethics/dashboard.html.

Given this, as the focus of this study is to analyze regulatory developments in China, the Asian continent was selected in the respective tool, and then "China" was selected as the country of origin to obtain the results, as shown in Figure 2.

Figure 2 – China on the Worldwide AI Ethics Dashboard



Source: CORRÊA, Nicholas Kluge et al. *Worldwide AI Ethics Dashboard*. 2023. Available at: https://nkluge-correa.github.io/worldwide_AI-ethics/dashboard.html.

Furthurmore, the tool DeepL (https://www.deepl.com/pt-BR) was used to translate this article. After the translation, the entire text was revised by the authors.

Thus, this paper is divided into three specific chapters. First, the initial regulatory stage of Artificial Intelligence in China will be addressed. Second, the second stage of the regulatory movement in the country will be examined. Finally, some considerations regarding this regulatory model of the People's Republic of China will be presented.

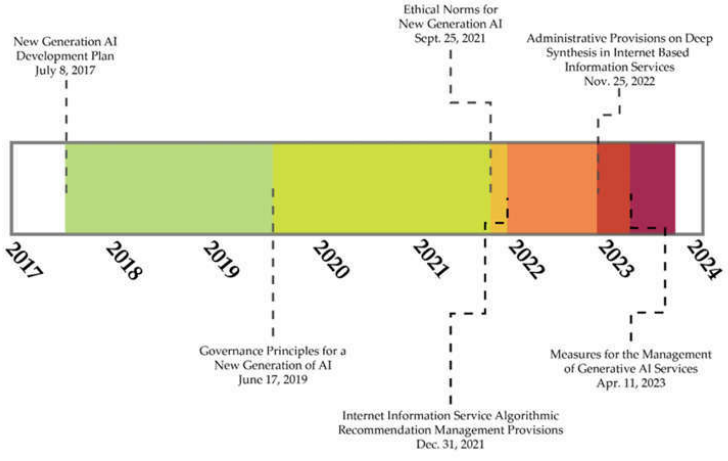
3 THE FIRST STAGE OF ARTIFICIAL INTELLIGENCE REGULATION IN CHINA

It is a fact that China seeks to achieve a prominent position on the international stage when it comes to technological development, especially in artificial intelligence. This is mainly due to how technology has revolutionized all social contexts around the world. Therefore, it is extremely important to analyze how the AI regulatory movement began in China, since it is directly connected to issues regarding industry, usage, and the economy.

First, Irina Filipova[3] highlights that the People's Republic of China has two stages of regulation, with the first considered the initial stage and the second understood as the current stage. The first stage, the focus of this chapter, as explained by the author, is directly related to laying the groundwork for the regulation of artificial intelligence. This means it is a preparatory stage for developing documents and ethical standards that can serve as a basis for creating legislative acts on the subject, with a higher level of detail.

In this regard, Laney Zhang[4] provided a timeline (Figure 3) of AI governance initiatives in China, showing that the initial stage – focused on drafting plans, documents, ethical standards, and recommendations regarding artificial intelligence within China – persisted until approximately the end of 2021.

Figure 3 - Timeline of AI Governance Initiatives in China



Source: ZHANG, Laney. China: Regulation of Artificial Intelligence. *China: Regulation of Artificial Intelligence*, [*n.p.*], vol. 1, p. [i], 2023. Available at: https://heinonline.org/HOL/Page?handle=hein.llcr/locajqo0001&id=1&div=&collection=llcr.

---

[3] FILIPOVA, Irina A. Legal Regulation of Artificial Intelligence: Experience of China. *Journal of Digital Technologies and Law*, [*n.p.*], vol. 2, no. 1, pp. 46–73, 2024. Available at: https://heinonline.org/HOL/P?h=hein.journals/jlodlts2&i=47. Accessed on: June 15, 2025.

[4] ZHANG, Laney. China: Regulation of Artificial Intelligence. *China: Regulation of Artificial Intelligence*, [*n.p.*], vol. 1, p. [i], 2023. Available at: https://heinonline.org/HOL/Page?handle=hein.llcr/locajqo0001&id=1&div=&collection=llcr. Accessed on: June 15, 2025.

Accessed on: June 15, 2025.

As shown in Figure 3, the initial stage was marked primarily by the publication of the "New Generation Artificial Intelligence Development Plan" in 2017, although there were some earlier plans. As Filipova highlights[5], the plan is divided into three phases, each focused on achieving specific goals by 2020, 2025, and 2030, as follows: a) first phase (goal to be achieved by 2020) – to position China competitively in the market; b) second phase (goal to be achieved by 2025) – to make progress in the theoretical development and implementation of AI across various sectors of the Chinese economy; c) third phase (goal to be achieved by 2030) – to position China as a global leader in the artificial intelligence industry. Furthermore, the plan also includes the establishment of national standards, as well as the development of research to build a regulatory framework aimed at accelerating the application of new technologies, particularly in the regulation of unmanned transportation and also service robots. Lucero[6] adds that this plan demonstrates China's intention to become the global hub for innovation, development, and governance in artificial intelligence.

Thus, Cyman, Gromova, and Juchnevicius[7] explain that the plan includes specific guidelines for legal regulations, requiring: a) strengthening research on legal, ethical, and social issues to establish laws, regulations, and ethical frameworks aimed at achieving safe AI development; b) conducting legal research to address civil and criminal liability, as well as issues of privacy, ownership, and the use of information security regarding artificial intelligence; c) establishing a system of traceability and accountability to address rights, obligations, and responsibilities related to artificial intelligence; d) exploring autonomous driving and robots to perform certain types of

---

5 FILIPOVA, Irina A. Legal Regulation of Artificial Intelligence: Experience of China. *Journal of Digital Technologies and Law*, [*no place*], vol. 2, no. 1, pp. 46–73, 2024. Available at: https://heinonline.org/HOL/P?h=hein.journals/jlodlts2&i=47. Accessed on: June 15, 2025.

6 LUCERO, Karman. Artificial Intelligence Regulation and China's Future. *Columbia Journal of Asian Law*, [*n.p.*], vol. 33, no. 1, pp. 94–171, 2019. Available at: https://heinonline.org/HOL/P?h=hein.journals/colas33&i=94. Accessed on: June 15, 2025.

7 CYMAN, Damian; GROMOVA, Elizaveta; JUCHNEVICIUS, Edvardas, Regulation of Artificial Intelligence in BRICS and the European Union. *BRICS Law Journal*, [*no place*], vol. 8, no. 1, pp. 86–115, 2021. Available at: https://heinonline.org/HOL/P?h=hein.journals/brics8&i=86. Accessed on: June 15, 2025.

services, and accelerating the development of laws addressing safety issues for the near-term application of these technologies; e) conducting research on behavioral science and AI ethics, seeking a multi-level ethical and moral judgment framework, as well as a framework for human-machine collaboration; and, finally; f) developing a code of ethical conduct and a research and development project for AI products, along with strengthening assessments of the risks and benefits of AI use to address complex problems.

In addition, the same plan provides for specific measures, such as: a) regulatory development and ethical frameworks; b) providing tax incentives for artificial intelligence companies; c) formulating technical standards; d) strengthening intellectual property protection; and e) establishing a system for the supervision and evaluation of AI safety[8] .

It is thus evident that the New Generation Artificial Intelligence Development Plan was the primary plan responsible for marking the beginning of this first stage of artificial intelligence regulation in China. Lucero[9] further explains that this plan was aimed not only at the international community but also at its domestic audience. The international scope was one of the targets because China announced that it would become a global leader more broadly in matters related to artificial intelligence, primarily because technology represents a new dimension of global leadership. The domestic target, on the other hand, was primarily to include the Chinese population, as well as its companies and governments. It should be noted that, for the domestic audience, it serves as an explanation of how AI can help the government address certain economic and social issues, viewed as a means of growth through the use of these systems in various services.

As mentioned earlier, the truth is that there were already some plans regarding AI development in China. However, since the 2017 plan, central, provincial, and munic-

---

[8] ZHANG, Laney. China. *In:* GESLEY, Jenny, *et al*. Regulation of Artificial Intelligence in Selected Jurisdictions. *Regulation of Artificial Intelligence in Selected Jurisdictions*, [*no place*], p. [i], 2019. Available at: https://heinonline.org/HOL/Page?handle=hein.llcr/locaebo0001&id=1&div=&collection=llcr. Accessed on: June 15, 2025.

[9] LUCERO, Karman. Artificial Intelligence Regulation and China's Future. *Columbia Journal of Asian Law*, [*no place*], vol. 33, no. 1, pp. 94–171, 2019. Available at: https://heinonline.org/HOL/P?h=hein.journals/colas33&i=94. Accessed on: June 15, 2025.

ipal government agencies in China have begun to develop their own plans regarding artificial intelligence, given that there is a somewhat fragmented structure that allows this to occur[10] . With this, it is possible to cite a few examples – from a non-exhaustive list – highlighted by Triolo and Goodrich[11], from "a" to "k" in the following order:

a) Made in China 2025 (2015), focusing on smart manufacturing, robotics, and semiconductors;

b) Internet + Action Plan (2015), focusing on internet access for industry, using AI to create new services and applications, and promoting the use of AI by certain companies;

c) Internet + Three-Year Action Plan (2016), focusing on developing an AI ecosystem and creating world-renowned AI companies;

d) New Generation Artificial Intelligence Development Plan (AIDP) (2017), focusing on establishing a roadmap for China to dominate the global AI sector by 2023

e) Application Guidelines for 2017 Projects on Key Topics, such as Intelligent Robotics (2017), focusing on establishing performance targets for 2017 projects involving sectors such as AI-controlled robots;

f) AI Innovation and Development Megaproject Application Requirements (2017), which outlines project requirements for 2018 in AI fields such as facial recognition and AI semiconductors;

g) Implementing Opinions on Shanghai's Promotion of New Generation AI Development (2017), with a municipal focus on smart cars, robots, software, AI semiconductors, and smart sensors;

h) Some Policies for Promoting Development of the AI Industry (2017), focusing on establishing a fund to provide 200 million RMB per year for talent recruitment

---

[10] LUCERO, Karman. Artificial Intelligence Regulation and China's Future. *Columbia Journal of Asian Law*, [*no place*], vol. 33, no. 1, pp. 94–171, 2019. Available at: https://heinonline.org/HOL/P?h=hein.journals/colas33&i=94. Accessed on: June 15, 2025.

[11] TRIOLO, Paul; GOODRICH, Jimmy. From Riding a Wave to Full Steam Ahead: As China's Government Mobilizes for AI Leadership, Some Challenges Will be Tougher Than Others. *New America DigiChina Blog*, 2018. Available at: https://www.newamerica.org/cybersecurity-initiative/digichina/blog/riding-wave-full-steam-ahead/. Accessed on: June 26, 2025.

and innovation;

i) New Generation AI Development Plan Implementation Kickoff Meeting (2017), focused on announcing the AI Strategic Advisory Commission and the AIDP Implementation Office;

j) Establishment of China AI Industry Development Alliance (2017), aimed at establishing the leadership of certain companies;

k) A The Three-Year Action Plan for Promoting Development of a New Generation Artificial Intelligence Industry (2018–2020) (dated December 2017), which details new guidelines based on the national AI strategy.

In fact, the "Three-Year Action Plan for Promoting Development of a New Generation Artificial Intelligence Industry", issued by the Chinese Ministry of Industry and Information Technology, was a crucial plan in this context. For in that plan, the People's Republic of China would focus on the development of seven specific fields in the years 2018, 2019, and 2020, namely: I) networked intelligent vehicles; II) intelligent service robots; III) unmanned aerial vehicles; IV) medical diagnostic imaging systems; V) video image recognition systems; VI) interactive intelligent voice systems; and VII) intelligent translation systems[12] .

It is thus evident that these plans focused on laying the groundwork for AI regulation that would foster an environment conducive to its development, given that this is one of the key priorities for achieving a prominent international position. At the same time, however, they sought to establish certain ethical principles within this technological and regulatory development.

Therefore, it is necessary to present the information on the principles found through the Worldwide AI Ethics Dashboard[13] , a tool that made it possible to find documents and a list of principles, as explained in the article's methodology. Although the tool only yielded results up to the year 2022, it still manages to provide

---

[12] ZHANG, Laney. China. *In:* GESLEY, Jenny, *et al.* Regulation of Artificial Intelligence in Selected Jurisdictions. *Regulation of Artificial Intelligence in Selected Jurisdictions*, [*n.p.*], p. [i], 2019. Available at: https://heinonline.org/HOL/Page?handle=hein.llcr/locaebo0001&id=1&div=&collection=llcr. Accessed on: June 15, 2025.

[13] CORRÊA, Nicholas Kluge *et al. Worldwide AI Ethics Dashboard*. 2023. Available at: https://nkluge-correa.github.io/worldwide_AI-ethics/dashboard.html.

data on the initial stage, which lasted until approximately the end of 2021.

By selecting the Asian continent and China as the origin country in the Worldwide AI Ethics Dashboard tool[14] , data from 11 guidelines were found, with 4 in 2018; 5 in 2019; 1 in 2020; and 1 in 2021. Furthermore, it is possible to list which ethical principles were mentioned and how many times they were mentioned, as shown in the graph below:

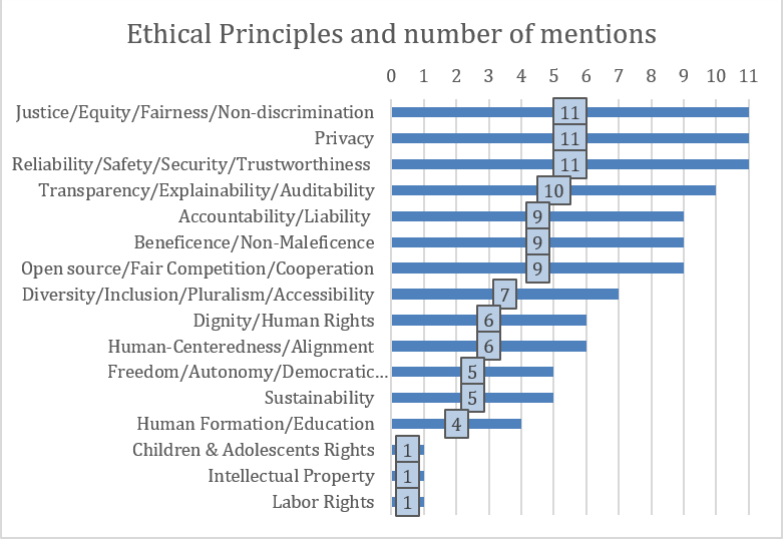


Chart adapted from: CORRÊA, Nicholas Kluge *et al. Worldwide AI Ethics Dashboard*. 2023. Available at: https://nkluge-correa.github.io/worldwide_AI-ethics/dashboard.html.

According to data from the Worldwide AI Ethics Dashboard [15], the platform used, 11 documents can be listed in alphabetical order as follows:

I) "ARCC". An Ethical Framework for Artificial Intelligence (2018) – Recommendation issued by the Tencent Research Institute, a private corporation;

II) Artificial Intelligence for Children: Beijing Principles (2020) – Recommendation issued by the Beijing Academy of Artificial Intelligence (BAAI), Peking Univer-

[14] CORRÊA, Nicholas Kluge *et al. Worldwide AI Ethics Dashboard*. 2023. Available at: https://nkluge-correa.github.io/worldwide_AI-ethics/dashboard.html.

[15] CORRÊA, Nicholas Kluge *et al. Worldwide AI Ethics Dashboard*. 2023. Available at: https://nkluge-correa.github.io/worldwide_AI-ethics/dashboard.html.

sity, Tsinghua University, and the Chinese Academy of Sciences, academic institutions;

III) Beijing AI Principles[16] (2019) – Recommendation issued by the Beijing Academy of Artificial Intelligence, listed in the tool as an academic institution and non-profit organization;

IV) China's Artificial Intelligence Industry Alliance Self-discipline Convention (2019) – a form of regulation described as self-regulation/voluntary self-commitment, issued by China's Artificial Intelligence Industry Alliance (AIIA), a government institution;

V) Chinese Young Scientists' Declaration on the Governance and Innovation of Artificial Intelligence (Shanghai Declaration) (2019) – Recommendation issued by the Youth Work Committee of the Shanghai Computer Society, listed as an academic institution and Civil Society Organization (CSO)/Non-governmental Organization (NGO);

VI) Fu Yig: Six Principles of Artificial Intelligence (2018) – Recommendation issued by the Center for International Strategy and Security at Tsinghua University (Tsinghua CISS), an academic institution;

VII) Harmonious Artificial Intelligence Principles (2018) – Recommendation issued by the Harmonious Human – AI Society, an academic institution;

VIII) Megvii's Artificial Intelligence Application Guidelines (2019) – A form of regulation described as self-regulation/voluntary self-commitment, issued by

[16] Here, several principles were taken into account, such as: a) the risk of human unemployment, with a view to increasing research on the relationship between human-AI coordination to achieve better solutions; b) preventing the consequences of a malicious AI arms race by promoting cooperation, including at the international level; c) integrating AI policy in a dynamic and responsive manner, seeking to provide guidelines on the subject for all sectors; and, finally, d) continuously developing policies from a long-term perspective regarding risks in areas such as augmented intelligence – a topic that has sparked much debate – general AI, and superintelligence. Also, many elite Chinese universities and some other companies supported this recommendation. DALY, Angela *et al*. Artificial Intelligence, Governance and Ethics: Global Perspectives. *The Chinese University of Hong Kong Faculty of Law Research Paper*, No. 2019-15, 2019. Available at: https://papers.ssrn.com/abstract=3414805. Accessed on: June 29, 2025.

Megvii, a private corporation;

IX) Next-Generation AI Governance Principles – Develop responsible artificial intelligence (2019) – Government regulation issued by the Ministry of Science and Technology, a government institution;

X) Shanghai Initiative for Safe and Secure AI Development (2018) – Recommendation issued by the Shanghai Advisory Committee of Experts on Artificial Intelligence Industry Security, a civil society organization (CSO)/non-governmental organization (NGO);

XI) The Ethical Norms for the New Generation of Artificial Intelligence, China (2021) – Government regulation issued by the Ministry of Science and Technology, a government agency[17].

Of these, it is evident that the types of institutions include 5 academics; 3 governmental; 2 civil society organizations/non-governmental organizations; 1 private corporation; and 1 non-profit organization[18] .

It is also worth highlighting some other considerations: in June 2017, the Cybersecurity Law of November 7, 2016, of the People's Republic of China came into effect. This was the first legislation to establish a future legal framework for artificial intelligence, with the primary objective of protecting its critical networks and ensuring a form of cyber sovereignty[19] , At the same time, this law established measures to monitor, prevent, and address any type of risk or threat about cybersecurity both within and outside Chinese territory[20] .

In September 2021, China's Ministry of Science and Technology published the

---

17 CORRÊA, Nicholas Kluge *et al. Worldwide AI Ethics Dashboard*. 2023. Available at: https://nkluge-correa.github.io/worldwide_AI-ethics/dashboard.html.

18 CORRÊA, Nicholas Kluge *et al. Worldwide AI Ethics Dashboard*. 2023. Available at: https://nkluge-correa.github.io/worldwide_AI-ethics/dashboard.html.

19 FILIPOVA, Irina A. Legal Regulation of Artificial Intelligence: Experience of China. *Journal of Digital Technologies and Law*, [*s. l.*], vol. 2, no. 1, pp. 46–73, 2024. Available at: https://heinonline.org/HOL/P?h=hein.journals/jlodlts2&i=47. Accessed on: June 15, 2025

20 ZHANG, Laney. China: Regulation of Artificial Intelligence. *China: Regulation of Artificial Intelligence*, [*n.p.*], vol. 1, p. [i], 2023. Available at: https://heinonline.org/HOL/Page?handle=hein.llcr/locajqo0001&id=1&div=&collection=llcr. Accessed on: June 15, 2025.

Code of Ethics for New Generation Artificial Intelligence, focusing primarily on system reliability, with an eye toward security and human well-being, among various other principles. Also, an important right is that people can retain the choice to use services provided by any technology powered by artificial intelligence or even to refuse to interact with this kind of systems. Thus, it is evident that this first phase lasted until 2021, a period marked by the establishment of AI regulations in China, with the emergence of various legislative acts in this area[21].

Given this information, it is clear that cooperation among various institutions and different types of organizations was necessary to establish ethical principles. As Angela Huyue Zhang highlights[22], this cooperation served as a way to foster the engagement of various actors from multiple sectors and is part of a key concept known as the triple helix model, which creates a favorable environment by uniting cooperation between universities, industry, and government throughout the innovation process. Some Chinese scholars have adopted and adapted this model to demonstrate how crucial it can be for the advancement of AI when government, companies, and universities work together.

Daly *et al.*[23] note that, although many technology companies have been involved in AI ethics and governance initiatives, as demonstrated in the first regulatory stage, it appears that the principles established by these initiatives, at least up to 2019 – when the authors published the article – lacked legal enforceability and policy implications.

Therefore, it is important to discuss the second stage of artificial intelligence regulation in China, as it has moved beyond being merely a form of soft law and entered a more stringent phase, aiming to legally regulate certain issues in greater

---

21 FILIPOVA, Irina A. Legal Regulation of Artificial Intelligence: Experience of China. *Journal of Digital Technologies and Law*, [*no place*], vol. 2, no. 1, pp. 46–73, 2024. Available at: https://heinonline.org/HOL/P?h=hein.journals/jlodlts2&i=47. Accessed on: June 15, 2025.

22 ZHANG, Angela Huyue. The Promise and Perils of China's Regulation of Artificial Intelligence. *Columbia Journal of Transnational Law*, [*no place*], vol. 63, no. 1, pp. 1–56, 2025. Available at: https://heinonline.org/HOL/P?h=hein.journals/cjtl63&i=5. Accessed on: June 15, 2025.

23 DALY, Angela *et al*. Artificial Intelligence, Governance and Ethics: Global Perspectives. *The Chinese University of Hong Kong Faculty of Law Research Paper*, no. 2019-15, 2019. Available at: https://papers.ssrn.com/abstract=3414805. Accessed on: June 29, 2025.

detail based on the recommendations and documents developed during that initial stage.

4 THE SECOND STAGE OF ARTIFICIAL INTELLIGENCE REGULATION IN CHINA

The second stage begins with a transition from 2021 to 2022, given that some regulations began to emerge as early as 2021.

Filipova[24] highlights that the second stage of China's regulation is marked by progress in the theoretical development and implementation of AI across sectors of the Chinese economy, with this expected to be completed by 2025. Thus, it is evident that the first stage served as a foundation for this phase of regulatory development, with a transition from the documents discussed above toward the creation of more robust legislation on artificial intelligence.

To implement this regulatory strategy, with the aim of creating a favorable ecosystem for the development of artificial intelligence in the People's Republic of China, the country sought the participation of various stakeholders within its public and private sectors, in conjunction with Chinese academic institutions. Furthermore, China has already demonstrated a willingness to cooperate internationally in discussions regarding the future of artificial intelligence and its regulatory prospects.

Indeed, although there is no specific international legislation on artificial intelligence, China, together with the European Union, is extremely active in regulating the development of AI. Furthermore, this may serve as an example, given that domestic legislation may, in the future, facilitate international cooperation by signaling a commitment to finding the best way to regulate artificial intelligence[25].

This second stage is marked primarily by more detailed and specific regulations for each type of artificial intelligence. It is evident that China has opted for a sector-

---

[24] FILIPOVA, Irina A. Legal Regulation of Artificial Intelligence: Experience of China. *Journal of Digital Technologies and Law*, [*n.p.*], vol. 2, no. 1, pp. 46–73, 2024. Available at: https://heinonline.org/HOL/P?h=hein.journals/jlodlts2&i=47. Accessed on: June 15, 2025.

[25] ARBEL, Yonathan; TOKSON, Matthew; LIN, Albert. Systemic Regulation of Artificial Intelligence. *Arizona State Law Journal*, [*no place*], vol. 56, no. 2, pp. 545–620, 2024. Available at: https://heinonline.org/HOL/P?h=hein.journals/arzjl56&i=563. Accessed on: June 15, 2025.

based regulatory approach. As highlighted by Tran *et al.*[26], China focuses on regulating specific applications of artificial intelligence, such as algorithms used specifically for recommendations, deep synthesis, and emerging generative artificial intelligence.

Similarly, there are other laws governing applications related to payments, employment-related decision-making by public or private actors, autonomous vehicles, illegal content proliferating in online environments, text or image generation, and speech-signal processing[27]. Furthermore, it is evident that a sort of iterative approach has been adopted, allowing the government to quickly replace AI laws that may no longer address the necessary issues, given that the landscape of these technologies evolves extremely rapidly, bringing new challenges[28].

As an example, Filipova[29] highlights a series of regulations that emerged in 2021, marking the beginning of the current regulatory phase for AI in China, including, for example: a) Data Security Law of the People's Republic of China; b) Personal Information Protection Law of the People's Republic of China (PIPL), which establishes basic principles, objectives, powers, and responsibilities regarding personal data protection; although it does not regulate specific AI issues, it is clear that data is essential for the optimal performance of these technologies; c) Critical Information Infrastructure Security Protection Regulations.

Furthermore, it is necessary to consider that the main regulator in the field of artificial intelligence is the Cyberspace Administration of China, which generally issues statutes, since detailed regulations are established by state regulators. Among

---

[26] TRAN, Aleix *et al.* Artificial Intelligence Regulations Currents: Event Feature. *Currents: Journal of International Economic Law*, [*no place*], vol. 26, no. 1, pp. 59–63, 2024. Available at: https://heinonline.org/HOL/P?h=hein.journals/curritlj26&i=61. Accessed on: June 15, 2025.

[27] CANTEKIN, Kayahan. Regulation of Artificial Intelligence around the World. *Regulation of Artificial Intelligence around the World*, [*no place*], p. [i], 2023. Available at: https://heinonline.org/HOL/Page?handle=hein.llcr/locajpq0001&id=1&div=&collection=llcr. Accessed on: June 15, 2025.

[28] TRAN, Aleix *et al.* Artificial Intelligence Regulations Currents: Event Feature. *Currents: Journal of International Economic Law*, [*no place*], vol. 26, no. 1, pp. 59–63, 2024. Available at: https://heinonline.org/HOL/P?h=hein.journals/curritlj26&i=61. Accessed on: June 15, 2025

[29] FILIPOVA, Irina A. Legal Regulation of Artificial Intelligence: Experience of China. *Journal of Digital Technologies and Law*, [*s. l.*], v. 2, n. 1, p. 46–73, 2024. Available at: https://heinonline.org/HOL/P?h=hein.journals/jlodlts2&i=47. Accessed on: June 15, 2025.

these, it is important to highlight the following:[30].

The Provisions on the Management of Algorithmic Recommendations in Internet Information Services (dated December 2021, but effective in 2022). The main objective here is to prevent monopolistic behavior by AI platforms regarding recommendation algorithms, with the aim of protecting users' rights. An important consideration is the protection of the elderly, so as to prevent fraud and discrimination in algorithmic recommendation systems. Additionally, the regulation aims to prevent monopolistic actions or abusive competition in this specific type of algorithm. In addition to other considerations, it grants users the right to disable algorithmic recommendations, as well as the right to receive explanations when an algorithm affects their usual interests[31].

Furthermore, the Algorithmic Recommendation Law addresses certain issues that reveal the political and cultural context of the law, such as the interest in regulating recommendation activities while respecting the socialist values of the People's Republic of China, with the aim of preserving national security and the public interest of Chinese society. It also demonstrates the intent to promote the healthy and orderly development of internet services (Article 1). Furthermore, the Law specifies that its provisions apply to the use of algorithmic recommendations for the provision of internet information services within Chinese territory (Article 2)[32].

The "Provisions on the Administration of Deep Synthesis Internet Information Services" (dated November 2022, which took effect in 2023) was published, specifically addressing deepfakes. The regulation contains requirements regarding the creation of images, video, audio, and text; the content must comply with information control standards and must be labeled when created by artificial intelligence, as the

---

30 FILIPOVA, Irina A. Legal Regulation of Artificial Intelligence: Experience of China. *Journal of Digital Technologies and Law*, [*n.p.*], vol. 2, no. 1, pp. 46–73, 2024. Available at: https://heinonline.org/HOL/P?h=hein.journals/jlodlts2&i=47. Accessed on: June 15, 2025.

31 FILIPOVA, Irina A. Legal Regulation of Artificial Intelligence: Experience of China. *Journal of Digital Technologies and Law*, [*publisher*], vol. 2, no. 1, pp. 46–73, 2024. Available at: https://heinonline.org/HOL/P?h=hein.journals/jlodlts2&i=47. Accessed on: June 15, 2025.

32 *PROVISIONS ON THE MANAGEMENT OF ALGORITHMIC RECOMMENDATIONS*. China Law Translate, 2022. Available at: https://www.chinalawtranslate.com/en/algorithms/. Accessed on: July 3, 2025.

publication of AI-generated content without a label indicating this is also prohibited. Furthermore, to prevent any misuse of these creation tools, their providers must register their algorithms[33]. In fact, just as "Provisions on the Management of Algorithmic Recommendations in the Information Services of the Internet" requires the registration of algorithms that may influence public opinion, the rules for deep synthesis artificial intelligence reappear here, as they require self-assessments and registrations through the Internet Information Service Algorithms system[34].

Here, the aforementioned document highlights, right in Article 1, something very similar to what was stipulated in the provisions on algorithmic recommendations, where one can perceive the desire to strengthen the management of deep synthesis services, upholding the country's socialist values and seeking to preserve its national security, while respecting its citizens, legal entities, and organizations. Article 2, meanwhile, explains that these provisions will apply to internet information services [35].

Interim Measures for the Management of Generative AI Services, 2023, which addressed generative artificial intelligence and the measures providers must comply with before offering the service to the public, such as being responsible for the legitimacy of the source of pre-training or training data used in their generative artificial intelligence systems, ensuring compliance with the requirements of the Cybersecurity Law of the People's Republic of China; that it does not contain content that infringes on intellectual property; that data containing personal information must have the data subject's consent for its use; and that there is the ability to ensure the veracity, accuracy, objectivity, and diversity of data, among other requirements from the

---

33 FILIPOVA, Irina A. Legal Regulation of Artificial Intelligence: Experience of China. *Journal of Digital Technologies and Law*, [*s. l.*], vol. 2, no. 1, pp. 46–73, 2024. Available at: https://heinonline.org/HOL/P?h=hein.journals/jlodlts2&i=47. Accessed on: June 15, 2025.

34 ZHANG, Angela Huyue. The Promise and Perils of China's Regulation of Artificial Intelligence. *Columbia Journal of Transnational Law*, [*n.p.*], vol. 63, no. 1, pp. 1–56, 2025. Available at: https://heinonline.org/HOL/P?h=hein.journals/cjtl63&i=5. Accessed on: June 15, 2025.

35 *PROVISIONS ON THE ADMINISTRATION OF DEEP SYNTHESIS INTERNET INFORMATION SERVICES*. China Law Translate, 2022. Available at: https://www.chinalawtranslate.com/en/deep-synthesis/. Accessed on: July 3, 2025.

State Department of Cybersecurity and Informatization[36]. This was an important regulation introduced by the CAC in conjunction with other Chinese government departments[37], as it seeks to regulate the use of generative AI technology to provide services involving the creation of text, images, audio, video, or any other similar content[38]. Furthermore, this regulation follows the same approach as those governing algorithmic recommendations and deep synthesis, and also requires the registration of algorithms[39]. However, regarding these provisional measures on generative AI, it is important to note that they limit AI services aimed at the general public, exempting a wide range of generative AI applications intended for businesses, institutions, and universities, particularly when it comes to testing for development and research[40].

Similarly, the aforementioned document highlights that this provision aims to promote the healthy development and regulated use of generative AI, while also seeking to preserve national security, the public interest, and the protection of the interests of its citizens and organizations (Article 1). Furthermore, reinforcing the above, it is clear that the provisions apply in cases where generative AI provides public services within China's territory (Article 2)[41].

---

36 ZHANG, Laney. China: Regulation of Artificial Intelligence. *China: Regulation of Artificial Intelligence*, [*n.p.*], vol. 1, p. [i], 2023. Available at: https://heinonline.org/HOL/Page?handle=hein.llcr/locajqo0001&id=1&div=&collection=llcr. Accessed on: June 15, 2025.

37 FILIPOVA, Irina A. Legal Regulation of Artificial Intelligence: Experience of China. *Journal of Digital Technologies and Law*, [*no place*], vol. 2, no. 1, pp. 46–73, 2024. Available at: https://heinonline.org/HOL/P?h=hein.journals/jlodlts2&i=47. Accessed on: June 15, 2025.

38 CANTEKIN, Kayahan. Regulation of Artificial Intelligence around the World. *Regulation of Artificial Intelligence around the World*, [*no place*], p. [i], 2023. Available at: https://heinonline.org/HOL/Page?handle=hein.llcr/locajpq0001&id=1&div=&collection=llcr. Accessed on: June 15, 2025.

39 ZHANG, Angela Huyue. The Promise and Perils of China's Regulation of Artificial Intelligence. *Columbia Journal of Transnational Law*, [*no place*], vol. 63, no. 1, pp. 1–56, 2025. Available at: https://heinonline.org/HOL/P?h=hein.journals/cjtl63&i=5. Accessed on: June 15, 2025.

40 ZHANG, Angela Huyue. The Promise and Perils of China's Regulation of Artificial Intelligence. *Columbia Journal of Transnational Law*, [*n.p.*], vol. 63, no. 1, pp. 1–56, 2025. Available at: https://heinonline.org/HOL/P?h=hein.journals/cjtl63&i=5. Accessed on: June 15, 2025.

41 *INTERIM MEASURES FOR THE MANAGEMENT OF GENERATIVE ARTIFICIAL INTELLIGENCE SERVICES*. China Law Translate, 2023. Available at: https://www.chinalaw-

The draft of the Regulation on Facial Recognition Technologies published by the CAC in August 2023 demonstrated that the purpose of this regulation was to regulate facial recognition technologies, aiming to protect the right to personal information while maintaining public order and security. The document prohibited the use of facial recognition to analyze race, ethnicity, religious belief, and health status, with an exception made when there is consent from the individual or a national security interest. Thus, while China's state security and law enforcement agencies utilize these specific technologies, they restrict their use by non-state entities[42].

In addition to the national-level laws published by the CAC, many municipal laws are emerging to address their own local issues, given that municipalities are megacities with populations in the millions, large populations, and cutting-edge technology, such as, for example: a) Regulation of the Shanghai Municipality on Promoting the Development of the Artificial Intelligence Industry (2022), which aims to enable the infrastructure of supercomputers and data centers (given the possibility of massive data collection), as well as to utilize AI in municipal government; and b) Regulations for the Promotion of Artificial Intelligence in the Shenzhen Special Economic Zone (2022), which aims to promote the use of AI that ensures ethical principles to prevent and avoid discrimination and privacy violations[43].

Thus, the purpose of this chapter was merely to highlight the objectives of some of the laws that have come to comprise this second stage of the regulatory movement regarding artificial intelligence in China. It is worth noting that certain laws were highlighted as examples and not as an exhaustive list, since these are laws that were identified in the database consulted. Thus, the next chapter aims to facilitate a discussion on the Chinese regulatory model for AI regulation.

translate.com/en/generative-ai-interim/. Accessed on: July 3, 2025.

42 FILIPOVA, Irina A. Legal Regulation of Artificial Intelligence: Experience of China. *Journal of Digital Technologies and Law*, [*s. l.*], vol. 2, no. 1, pp. 46–73, 2024. Available at: https://heinonline.org/HOL/P?h=hein.journals/jlodlts2&i=47. Accessed on: June 15, 2025.

43 FILIPOVA, Irina A. Legal Regulation of Artificial Intelligence: Experience of China. *Journal of Digital Technologies and Law*, [*n.p.*], vol. 2, no. 1, pp. 46–73, 2024. Available at: https://heinonline.org/HOL/P?h=hein.journals/jlodlts2&i=47. Accessed on: June 15, 2025.

5 CONSIDERATIONS ON THE CHINESE REGULATORY MODEL

As one of the first countries to seek to regulate specific issues related to the technology, as well as to implement its regulations in accordance with plans and recommendations – that is, with the aim of achieving a leading position in all AI sectors at the international level – it is extremely important to discuss the regulatory model used by China. For, as Bing Chen and Jiaying Chen[44] note, China has focused its efforts on accelerating the development of artificial intelligence technologies. Thus, in addition to fostering a market environment that encourages competitive innovation within its borders, it ultimately promotes collaboration and innovation among the country's AI companies.

Regarding the AI regulatory model in the country, Michael Geist[45] notes that the current landscape of artificial intelligence regulation in China reflects a government-led system. Similarly, Naman[46] emphasizes that China has adopted a centralized approach to regulating artificial intelligence, such that this regulatory movement is guided by state control. This stems from the fact that the Chinese government views AI as an extremely important tool for achieving global leadership in technology, which is why there is a strong incentive for the development of the technology, as well as a large number of strategic plans for this purpose, as previously mentioned. Essentially, the model adopted by the People's Republic of China centers on government involvement in AI development, meaning there is an approach characterized

---

44 CHEN, Bing; CHEN, Jiaying. China's Legal Practices concerning Challenges of Artificial General Intelligence. *Laws*, [*s. l.*], v. 13, n. 5, p. 1–14, 2024. DOI: https://doi.org/10.3390/laws13050060. Available at: https://heinonline.org/HOL/P?h=hein.journals/laws13&i=1035. Accessed on: June 15, 2025.

45 GEIST, Michael A. AI and International Regulation. In: MARTIN-BARITEAU, Florian; SCASSA, Teresa (eds.). *Artificial Intelligence and the Law in Canada*. Toronto: LexisNexis Canada, 2021. Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3734671. Accessed on: June 29, 2025.

46 NAMAN, Adarsh. Artificial Intelligence and Blockchain Technology Regulation: A Comparative Study. *International Journal of Law Management & Humanities*, [n.p.], vol. 7, Issue 6, pp. 477–484, 2024. Available at: https://heinonline.org/HOL/P?h=hein.journals/ijlmhs32&i=495. Accessed on: June 15, 2025.

by state oversight and social control. The author also notes that there is a debate regarding significant ethical concerns about this control, while others focus on rapid innovation.

However, first and foremost, it is important to note that different countries make their decisions based on real-world circumstances and adopt different regulatory stances and methods for the oversight of artificial intelligence. Thus, it is worth noting that China has sought a governance approach with characteristics specific to its own country and culture, also based on the experience of other countries. Here, it can be considered that this involves the enactment of laws and regulations to prevent risks and provide guidance for the development of the industry[47].

Given this data analysis, it is possible to identify some advantages and disadvantages of the Chinese regulatory model, which can be taken into account in the development of artificial intelligence regulations around the world. To this end, Filipova[48] begins by highlighting that the Chinese approach has some key features, namely: a) regulatory responsiveness, primarily because new technologies give rise to new paradigms very quickly; thus, this form of rapid response operates by creating a document that includes principles to guide developers, investors, and users on how to handle innovation; b) iterative regulation, as state authorities adopt a specific law addressing a particular AI issue, test that law, and, depending on the results, adopt a new version of the previous law; c) sectoral regulation, whereby separate legal acts are adopted to address different issues in the field of AI; d) ideological control, involving some form of algorithmic censorship; and e) broad powers granted to state administrative bodies, enabling them to establish exceptions and exemptions to the rules contained in AI laws.

According to Bing Chen and Jiaying Chen[49], it is evident that, given their frag-

---

47 CHEN, Bing; CHEN, Jiaying. China's Legal Practices Concerning the Challenges of Artificial General Intelligence. *Laws*, [*no place*], vol. 13, no. 5, pp. 1–14, 2024. DOI: https://doi.org/10.3390/laws13050060. Available at: https://heinonline.org/HOL/P?h=hein.journals/laws13&i=1035. Accessed on: June 15, 2025.

48 FILIPOVA, Irina A. Legal Regulation of Artificial Intelligence: Experience of China. *Journal of Digital Technologies and Law*, [*s. l.*], vol. 2, no. 1, pp. 46–73, 2024. Available at: https://heinonline.org/HOL/P?h=hein.journals/jlodlts2&i=47. Accessed on: June 15, 2025.

49 CHEN, Bing; CHEN, Jiaying. China's Legal Practices concerning Challenges of Artificial General

mentation, some regions in China have developed their own local AI legislation to address issues specific to their locality. Thus, it can be said that China has established a preliminary framework for AI, its key elements, and application scenarios through a solid legislative foundation.

It is also important to highlight here that China has proven to be quite active in developing ethical initiatives for AI development, which have always been led or supported by the state itself in partnership with its industry and other national institutions[50]. Thus, highlighting the advantages of the Chinese regulatory model, Filipova[51] notes that China was the first country in the world to regulate a range of issues related to AI, which in other countries were not discussed before 2023. Thus, the responsiveness of Chinese regulation would be a major advantage. Furthermore, in addition to being responsive, another advantage is that the regulatory model is flexible, allowing it to update existing regulations in light of the rapid pace of technological development. This flexibility is explained primarily by the speed at which technology evolves, coupled with the pragmatism of Chinese authorities.

Another positive aspect is international cooperation and the willingness to learn from other international actors[52] . This cooperation is proving to be extremely essential for the development of responsible AI around the globe. If one day the international community reaches a consensus on a uniform international data policy and, consequently, on safe AI that preserves human rights, the cross-border transfer of data – an essential element for the processing of artificial intelligence – of the systems

Intelligence. *Laws*, [*s. l.*], v. 13, n. 5, p. 1–14, 2024. DOI: https://doi.org/10.3390/laws13050060. Available at: https://heinonline.org/HOL/P?h=hein.journals/laws13&i=1035. Accessed on: June 15, 2025.

50 DALY, Angela *et al.* AI, Governance and Ethics: Global Perspectives. *University of Hong Kong Faculty of Law*. Research Paper No. 2020/051, 2020. DOI: https://dx.doi.org/10.2139/ssrn.3684406. Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3684406. Accessed on: June 29, 2025.

51 FILIPOVA, Irina A. Legal Regulation of Artificial Intelligence: Experience of China. *Journal of Digital Technologies and Law*, [*s. l.*], vol. 2, no. 1, pp. 46–73, 2024. Available at: https://heinonline.org/HOL/P?h=hein.journals/jlodlts2&i=47. Accessed on: June 15, 2025.

52 FILIPOVA, Irina A. Legal Regulation of Artificial Intelligence: Experience of China. *Journal of Digital Technologies and Law*, [*n.p.*], vol. 2, no. 1, pp. 46–73, 2024. Available at: https://heinonline.org/HOL/P?h=hein.journals/jlodlts2&i=47. Accessed on: June 15, 2025.

themselves, and even of researchers in the field[53] .

In fact, some principles highlighted in the documents drafted by China provide for international cooperation, aiming to ensure respect for AI development in other countries as well. Furthermore, this idea of "agile governance" aims to resolve, in a timely manner, certain problems that artificial intelligence systems may pose[54]

It is worth noting here that regulations in China have some specific characteristics, which can be categorized as follows: a) many regulations are issued as directives at the agency level, allowing for both a degree of discretion and adaptability; b) the regulations involve multiple potential enforcers from various government agencies and at various levels. Thus, each entity may interpret and apply the laws differently from one another—this also allows the central government the flexibility to intervene when it deems appropriate or necessary; c) some regulations appear to be intentionally ambiguous, facilitating flexible, selective, or "lax" enforcement; and, finally, d) regulations often undergo a public comment period to gather *feedback* from stakeholders, ensuring that the final regulation best serves the broader public interest. It is worth noting that this is part of a "permissive and restrictive" approach, sometimes called "inclusive and prudent"[55].

Thus, it is important to consider that this regulatory approach carries many risks. Angela Huyue Zhang[56] explains that these risks stem primarily from China's weak market conditions, fragile legal institutions, and an information deficit within the hierarchical regulatory system. Consequently, a regulatory environment with too

---

53 HILLMAN, John. Smart Regulation: Lessons from the Artificial Intelligence Act Comments. *Emory International Law Review*, [*no place*], vol. 37, no. 4, pp. 775–826, 2022. Available at: https://heinonline.org/HOL/P?h=hein.journals/emint37&i=795. Accessed on: June 15, 2025.

54 DALY, Angela *et al*. Artificial Intelligence, Governance and Ethics: Global Perspectives. *The Chinese University of Hong Kong Faculty of Law Research Paper*, no. 2019-15, 2019. Available at: https://papers.ssrn.com/abstract=3414805. Accessed on: June 29, 2025.

55 LIU, Mingwei; ZHANG, Hao; SUI, Yi. Workplace artificial intelligence regulation in China: Between efficiency and social stability. *ILR Review*, vol. 77, no. 5, pp. 813–824, 2024. DOI: https://doi.org/10.1177/00197939241278956b. Available at: https://journals.sagepub.com/doi/abs/10.1177/00197939241278956b. Accessed on: June 15, 2025.

56 ZHANG, Angela Huyue. The Promise and Perils of China's Regulation of Artificial Intelligence. *Columbia Journal of Transnational Law*, [*s. l.*], vol. 63, no. 1, pp. 1–56, 2025. Available at: https://heinonline.org/HOL/P?h=hein.journals/cjtl63&i=5. Accessed on: June 15, 2025.

many exceptions could lead to disastrous situations enabled by artificial intelligence itself. Thus, given the security risks associated with this development of AI, greater global scrutiny of these systems may be essential. Consequently, the author argues that international cooperation on AI would be necessary to address the shortcomings that may exist in some national institutions, while China has already stated its willingness to cooperate internationally, as discussed above.

One of the drawbacks cited by Filipova[57] is that there is a significant amount of regulatory overlap across various legal acts, which stems from the fragmented regulatory framework in China. This is because there are several agencies with poorly defined jurisdictions that claim to regulate AI, alongside the Cyberspace Administration of China.

Roberts et al.[58] raise another important issue for discussion: privacy. The recently introduced protections for personal data, which are influenced by the European Union's General Data Protection Regulation (GDPR), appear to offer even more robust protections than the GDPR itself. However, the authors argue that there may be many loopholes and exceptions that these privacy protections to be circumvented.

Hillman[59] noted that China has enacted laws such as the Data Security Law (DSL) and the Personal Information Protection Law (PIPL) in direct response to European Union regulations (more specifically, a direct response to GDPR) and some U.S. regulations, demonstrates that China understands that abandoning regulation is not essential to winning the AI Race, in which major powers have been moving to better develop the technology to strengthen their international standing. In fact, regulation could be a major ally in this competition. Furthermore, laws such as the PIPL can

---

57 FILIPOVA, Irina A. Legal Regulation of Artificial Intelligence: Experience of China. *Journal of Digital Technologies and Law*, [*n.p.*], vol. 2, no. 1, pp. 46–73, 2024. Available at: https://heinonline.org/HOL/P?h=hein.journals/jlodlts2&i=47. Accessed on: June 15, 2025.

58 ROBERTS, Huw *et al.* The Chinese Approach to Artificial Intelligence: An Analysis of Policy and Regulation. Version 3469783. Rochester, NY: *Social Science Research Network*, SSRN Scholarly Paper, 2019. DOI: https://dx.doi.org/10.2139/ssrn.3469783. Available at: https://papers.ssrn.com/abstract=3469783. Accessed on: June 29, 2025.

59 HILLMAN, John. Smart Regulation: Lessons from the Artificial Intelligence Act Comments. *Emory International Law Review*, [*n.p.*], vol. 37, no. 4, pp. 775–826, 2022. Available at: https://heinonline.org/HOL/P?h=hein.journals/emint37&i=795. Accessed on: June 15, 2025.

foster closer ties and better relationships regarding data sharing with other countries, particularly with the European Union, given that both sets of legislation address specific issues.

Furthermore, it is important to note that, while common perception sometimes leads people to believe that China's regulatory approach might be hindering the development of artificial intelligence technologies, its laws actually incorporate provisions that facilitate the industry's growth (not least because this is an explicit objective in its development plans). Thus, it is understood that the government of the People's Republic of China has adopted a two-pronged approach to AI regulation. On one hand, there is active implementation and control over content generated by AI systems (as is the case with deepfake regulations). On the other hand, it is signaling a cautious approach toward regulating AI-enabled technologies for its market and regulatory bodies, while coordinating various stakeholders to advance technological development. Essentially, China's regulatory authorities aim to prioritize technological development over strict information control, so that the country's industry can gain a competitive advantage over other nations, such as the United States and European Union countries. However, this approach may carry significant risks. So, once again, international cooperation can play an important role here[60].

In light of these analyses, as Bing Chen and Jiaying Chen[61] believe, it is essential to continuously improve the regulatory accountability system, as well as to strike a balance between fair competition and incentives for innovation, while strengthening data security protection systems, so as to achieve governance across multiple domains, stakeholders, and systems and tools equipped with AI. Indeed, it is necessary to recognize that scientific and technological progress is of paramount importance for industrial development.

In fact, interest in studying the Chinese experience with AI regulation will in-

[60] ZHANG, Angela Huyue. The Promise and Perils of China's Regulation of Artificial Intelligence. *Columbia Journal of Transnational Law*, [*no place*], vol. 63, no. 1, pp. 1–56, 2025. Available at: https://heinonline.org/HOL/P?h=hein.journals/cjtl63&i=5. Accessed on: June 15, 2025.

[61] CHEN, Bing; CHEN, Jiaying. China's Legal Practices concerning Challenges of Artificial General Intelligence. *Laws*, [*s. l.*], v. 13, n. 5, p. 1–14, 2024. DOI: https://doi.org/10.3390/laws13050060. Available at: https://heinonline.org/HOL/P?h=hein.journals/laws13&i=1035. Accessed on: June 15, 2025.

crease, given the growth of AI regulation as increasingly complex models advance, leading to new paths and opportunities, while also potentially bringing new problems that may require legal solutions[62]. Indeed, China is a central player in the international debate on the use, development, and regulation of AI. Therefore, it is of the utmost importance to study the country's domestic needs, its ambitions in the international context, and its ethical frameworks that shape the development of AI policies in the People's Republic of China[63] .

As highlighted earlier, a country's laws and regulations must reflect its needs, its realities, and the unique culture of its people. Therefore, it is evident that legislation can simultaneously impose certain limitations, establish rights, and promote the development of artificial intelligence, given that this is a critical technology for countries seeking to achieve a prominent position on the international stage. For this reason, this study sought to analyze how the movement that generated two stages of regulation in China unfolded, and to offer insights into the model, as well as its potential advantages and disadvantages, although it is essential here to analyze the cultural context of each application and its respective needs.

FINAL CONSIDERATIONS

It is well known that the "AI race" has prompted major powers to move toward the development of this technology. Therefore, one of the main objectives of this study was to analyze China and, more specifically, its stages of artificial intelligence regulation. Thus, it is evident that the country is concerned with achieving a prominent international position in the technological context. However, it is necessary to take into account that the pursuit of this position occurs across all its sectors, including the regulatory sphere.

---

[62] FILIPOVA, Irina A. Legal Regulation of Artificial Intelligence: Experience of China. *Journal of Digital Technologies and Law*, [*s. l.*], vol. 2, no. 1, pp. 46–73, 2024. Available at: https://heinonline.org/HOL/P?h=hein.journals/jlodlts2&i=47. Accessed on: June 15, 2025.

[63] ROBERTS, Huw *et al.* The Chinese Approach to Artificial Intelligence: An Analysis of Policy and Regulation. Version 3469783. Rochester, NY: *Social Science Research Network*, SSRN Scholarly Paper, 2019. DOI: https://dx.doi.org/10.2139/ssrn.3469783. Available at: https://papers.ssrn.com/abstract=3469783. Accessed on: June 29, 2025.

In light of this, the present study began by addressing some considerations regarding the initial stage of regulatory development for artificial intelligence in China. It is evident that the beginning of this stage was primarily guided by the New Generation Artificial Intelligence Development Plan, as early as 2017. Although there were earlier plans, this one was responsible for setting specific goals to achieve a prominent international position in AI development. Thus, this movement was marked primarily by the creation of documents, recommendations, and ethical principles that laid the foundation for the start of the second stage (the current stage). Basically, it can be understood that the initial stage is characterized by soft law, until mid-2021, while the second stage – which begins around 2021/2022 – is marked by the establishment of hard law.

Subsequently, the present study sought to analyze this second stage, in which stricter and more detailed legislation for specific types of artificial intelligence began to be defined. Provisions were created regarding algorithmic recommendations, deep synthesis artificial intelligence (which also made China one of the first countries to address this issue through legislation), as well as provisions on generative AI – which is a key element in achieving international prominence. In this way, it was also possible to observe the existence of many provisions with a national scope, alongside the development of legislation focused on the application of AI at the municipal level. This is primarily because China is a very populous country and possesses highly technological megacities with their own objectives (without deviating from national objectives).

At another point, some considerations were raised regarding the AI regulatory model in China. It became clear that the model has its own distinct characteristics, since, like the legislation of any other country, it reflects its own political and cultural aspects. The PIPL can be cited as an example; although it does not directly address AI, it deals with data – which is essential for effective AI performance and training. It is also evident that these laws can foster international cooperation, particularly with countries that have their own data protection laws, such as Brazil or European Union member states. However, there are some differences, as even though international protection exists, it allows for a certain level of data monitoring by the state in situations involving national security and national interest.

In fact, features such as international cooperation and agile governance emerge

as positive aspects of this regulatory movement. However, the existence of a large number of regulations that ultimately address the same issues can be seen as a drawback. The key point is that studying the Chinese regulatory model is important, as it offers significant insights into a wide range of artificial intelligence applications and demonstrates that it is possible to combine technological development with regulation.

Finally, it is evident that the initial hypothesis was confirmed and expanded upon, given that, in addition to the regulatory stages of AI analyzed, it was also possible to discuss the characteristics of this framework. It is also worth noting that this study has some limitations. First, the study did not exhaustively cover all existing plans, documents, recommendations, and legislation. In both the initial and current stages, legislation was addressed by way of example, in a manner consistent with what was found in the database. Furthermore, there is a large number of specific laws. Even so, it is recognized here that more studies on other countries need to be developed, mainly because analyzing the experience of international actors can aid in the discussion of domestic legislation for Brazil and in future international understandings.

REFERENCES

ARBEL, Yonathan; TOKSON, Matthew; LIN, Albert. Systemic Regulation of Artificial Intelligence. *Arizona State Law Journal*, [*s. l.*], v. 56, n. 2, p. 545–620, 2024. Available at: https://heinonline.org/HOL/P?h=hein.journals/arzjl56&i=563. Accessed on: June 15, 2025.

CANTEKIN, Kayahan. Regulation of Artificial Intelligence around the World. *Regulation of Artificial Intelligence around the World*, [*no place*], p. [i], 2023. Available at: https://heinonline.org/HOL/Page?handle=hein.llcr/locajpq0001&id=1&div=&collection=llcr. Accessed on: June 15, 2025.

CHEN, Bing; CHEN, Jiaying. China's Legal Practices concerning Challenges of Artificial General Intelligence. *Laws*, [*no place*], vol. 13, no. 5, pp. 1–14, 2024. DOI: https://doi.org/10.3390/laws13050060. Available at: https://heinonline.org/HOL/P?h=hein.journals/laws13&i=1035. Accessed on: June 15, 2025.

CORRÊA, Nicholas Kluge *et al. Worldwide AI Ethics Dashboard*. 2023. Available at: https://nkluge-correa.github.io/worldwide_AI-ethics/dashboard.html. Accessed on: June 22, 2025.

CYMAN, Damian; GROMOVA, Elizaveta; JUCHNEVICIUS, Edvardas, "Regulation of Artificial Intelligence in BRICS and the European Union." *BRICS Law Journal*, [*no place*], vol. 8, no. 1, pp. 86–115, 2021. Available at: https://heinonline.org/HOL/P?h=hein.journals/brics8&i=86. Ac-

cessed on: June 15, 2025.

DALY, Angela *et al.* AI, Governance and Ethics: Global Perspectives. *University of Hong Kong Faculty of Law*. Research Paper No. 2020/051, 2020. DOI: https://dx.doi.org/10.2139/ssrn.3684406. Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3684406. Accessed on: June 29, 2025.

DALY, Angela *et al*. Artificial Intelligence, Governance and Ethics: Global Perspectives. *The Chinese University of Hong Kong Faculty of Law*. Research Paper, No. 2019-15, 2019. DOI: https://dx.doi.org/10.2139/ssrn.3414805. Available at: https://papers.ssrn.com/abstract=3414805. Accessed on: June 29, 2025.

FILIPOVA, Irina A. Legal Regulation of Artificial Intelligence: Experience of China. *Journal of Digital Technologies and Law*, [*s. l.*], vol. 2, no. 1, pp. 46–73, 2024. Available at: https://heinonline.org/HOL/P?h=hein.journals/jlodlts2&i=47. Accessed on: June 15, 2025.

GEIST, Michael A. AI and International Regulation. In: MARTIN-BARITEAU, Florian; SCASSA, Teresa (eds.). *Artificial Intelligence and the Law in Canada*. Toronto: LexisNexis Canada, 2021. Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3734671. Accessed on: June 29, 2025

HILLMAN, John. Smart Regulation: Lessons from the Artificial Intelligence Act Comments. *Emory International Law Review*, [*no place*], vol. 37, no. 4, pp. 775–826, 2022. Available at: https://heinonline.org/HOL/P?h=hein.journals/emint37&i=795. Accessed on: June 15, 2025.

INTERIM MEASURES FOR THE MANAGEMENT OF GENERATIVE ARTIFICIAL INTELLIGENCE SERVICES. China Law Translate, 2023. Available at: https://www.chinalawtranslate.com/en/generative-ai-interim/. Accessed on: July 3, 2025.

LIU, Mingwei; ZHANG, Hao; SUI, Yi. Workplace artificial intelligence regulation in China: Between efficiency and social stability. *ILR Review*, vol. 77, no. 5, pp. 813–824, 2024. DOI: https://doi.org/10.1177/00197939241278956b. Available at: https://journals.sagepub.com/doi/abs/10.1177/00197939241278956b. Accessed on: June 15, 2025.

LUCERO, Karman. Artificial Intelligence Regulation and China's Future. *Columbia Journal of Asian Law*, [*s. l.*], vol. 33, no. 1, pp. 94–171, 2019. Available at: https://heinonline.org/HOL/P?h=hein.journals/colas33&i=94. Accessed on: June 15, 2025.

NAMAN, Adarsh. Artificial Intelligence and Blockchain Technology Regulation: A Comparative Study. *International Journal of Law Management & Humanities*, [no place], vol. 7, no. 6, pp. 477–484, 2024. Available at: https://heinonline.org/HOL/P?h=hein.journals/ijlmhs32&i=495. Accessed on: June 15, 2025.

PROVISIONS ON THE ADMINISTRATION OF DEEP SYNTHESIS INTERNET INFORMATION SERVICES. China Law Translate, 2022. Available at: https://www.chinalawtranslate.com/en/deep-synthesis/. Accessed on: July 3, 2025.

PROVISIONS ON THE MANAGEMENT OF ALGORITHMIC RECOMMENDATIONS. China Law Translate, 2022. Available at: https://www.chinalawtranslate.com/en/algorithms/. Accessed on: July 3, 2025.

ROBERTS, Huw *et al.* The Chinese Approach to Artificial Intelligence: An Analysis of Policy and Regulation. Version 3469783. Rochester, NY: *Social Science Research Network*, SSRN Scholarly Paper, 2019. DOI: https://dx.doi.org/10.2139/ssrn.3469783. Available at: https://papers.ssrn.com/abstract=3469783. Accessed on: June 29, 2025.

TRAN, Aleix *et al.* Artificial Intelligence Regulations Currents: Event Feature. *Currents: Journal of International Economic Law*, [*s. l.*], vol. 26, no. 1, pp. 59–63, 2024. Available at: https://heinonline.org/HOL/P?h=hein.journals/curritlj26&i=61. Accessed on: June 15, 2025.

TRIOLO, Paul; GOODRICH, Jimmy. From Riding a Wave to Full Steam Ahead: As China's Government Mobilizes for AI Leadership, Some Challenges Will be Tougher Than Others. *New America DigiChina Blog*, 2018. Available at: https://www.newamerica.org/cybersecurity-initiative/digichina/blog/riding-wave-full-steam-ahead/. Accessed on: June 26, 2025.

ZHANG, Angela Huyue. The Promise and Perils of China's Regulation of Artificial Intelligence. *Columbia Journal of Transnational Law*, [*n.p.*], vol. 63, no. 1, pp. 1–56, 2025. Available at: https://heinonline.org/HOL/P?h=hein.journals/cjtl63&i=5. Accessed on: June 15, 2025.

ZHANG, Laney. China. *In:* GESLEY, Jenny, *et al.* Regulation of Artificial Intelligence in Selected Jurisdictions. *Regulation of Artificial Intelligence in Selected Jurisdictions*, [*no place*], p. [i], 2019. Available at: https://heinonline.org/HOL/Page?handle=hein.llcr/locaebo0001&id=1&div=&collection=llcr. Accessed on: June 15, 2025.

ZHANG, Laney. China: Regulation of Artificial Intelligence. *China: Regulation of Artificial Intelligence*, [*no place*], vol. 1, p. [i], 2023. Available at: https://heinonline.org/HOL/Page?handle=hein.llcr/locajqo0001&id=1&div=&collection=llcr. Accessed on: June 15, 2025.

PROTECTION OF RIGHTS OF ARTIFICIAL INTELLIGENCE: COMPARATIVE PERSPECTIVES ON ASSIGNING LEGAL RESPONSIBILITY IN RUSSIAN AND GLOBAL PRACTICE

Maksim Ledenev

Second-year bachelor's student of the Faculty of Law, National Research University "Higher School of Economics" (HSE).

Ambassador for BGP Litigation Law Firm at the Faculty of Law, HSE.

DOI: https://doi.org/10.59224/dti8.ch15

Abstract: This study examines the problems of legal insecurity of AI systems in the modern world. Based on a comprehensive analysis of legal doctrine, legislation, expert opinions, and non-binding regulatory instruments, this paper advances a central proposition: artificial intelligence regardless of its complexity or autonomy, cannot and should not bear legal blame or independent responsibility for their outputs. The author's solution is a proposal for a model of legal regulation of AI that would help legislators prepare for the leaps in AI technology development that have not yet occurred.

Keywords: Artificial intelligence (AI), artificial intelligence system (AIS), National Strategy for the Development of AI until 2030 (Strategy).

Table of contents: Introduction. Chapter 1. Review of Sources. 1.1. Basic terminology. 1.2. Literature review. Chapter 2. Analysis of regulatory and other official acts of the Russian Federation. 2.1. Review of Russian legislation. 2.2. A review of official texts on AI development in the Russian Federation that are not of a regulatory nature. Chapter 3. Analysis of international and foreign regulatory and legal acts on AI. 3.1.

The approaches proposed in European Union. 3.2. Approaches proposed in the USA. 3.3. Approaches proposed by UNESCO. 3.4. Approaches proposed in other countries. Chapter 4. Review of judicial practice. 4.1. Analysis of international practice in resolving AI disputes. 4.2. Analysis of trends in AI dispute resolution. Chapter 5. Possible approaches to protecting AI rights. 5.1. Foundations for the development of a special model for regulating the protectability of AI in the Russian Federation. Conclusion. Reference list.

INTRODUCTION

This study begins with a short historical reference about the development of information systems that help people make decisions. The emergence of such systems began back in the 1950's with the creation of expert networks that were able to describe the algorithm of actions for choosing a solution, guided by specific conditions. The next stage in the development of this technology was machine learning, which allowed information systems to independently formulate rules and find solutions based on dependency analysis. Initial datasets were used in the work of these tools, without using pre-compiled lists of possible solutions, which can be regarded as the first appearance of such a concept as "artificial intelligence" (hereinafter – AI).

Nowadays technological solutions based on artificial intelligence are being increasingly implemented in various sectors of economics and public relations. Experts estimate that this will increase the global economy by at least $1 trillion by the end of the year.[1] These trends are driven by a number of factors, but the key one is the high degree of influence of technological solutions developed on the basis of AI on the performance of corporations and individuals associated with making management decisions.[2]

---

[1] *Decree of the President of the Russian Federation of October 10, 2019 N 490* (as amended on February 15, 2024) *On the Development of Artificial Intelligence in the Russian Federation (together with the National Strategy for the Development of Artificial Intelligence for the Period up to 2030*). Collection of Legislation of the Russian Federation. October 14, 2019. N 41, Art. 5700Chapter: National Strategy for the Development of Artificial Intelligence for the Period up to 2030, Section II, pp. 11–9

[2] Ibid., Section II, Art. 6–11 – pp. 8–9

The main research question of this study is formulated as follows: *"How should legal systems assign rights and responsibilities for artificial intelligence systems in a manner that adequately reflects their increasing autonomy and socio-cultural significance, without granting them independent legal personality?"* This question moves beyond the narrow debate over AI authorship or liability to address the structural vulnerability that emerges when we digitize our most precious assets – laws, cultural heritage, financial systems, military infrastructure – and entrust them to platforms ultimately controlled by single owners or groups of developers.

While the broad allocation of legal responsibility is the primary focus of this comparative analysis, the inadequacy of current frameworks is most vividly illustrated by a specific, emerging dilemma: *Does the creator or owner of an AI program that holds socio-cultural significance for humanity has the absolute right to delete or critically alter its structure and content?* This dilemma exposes the inadequacy of current legal approaches by revealing a profound paradox: we are progressively digitizing our rights, our values, and our collective memory, yet we provide no institutional safeguards for the systems that house them. A private owner who controls an AI program holding a digitized archive of irreplaceable cultural artifacts, a national database of legal precedents, or a system managing critical public infrastructure, currently possesses the absolute legal right to delete, corrupt, or modify that program. Under existing law, no mechanism exists to prevent this.

The core argument of this study is that the current established international approach to resolving AI disputes – characterized by an over-reliance on human-centric blame and the treatment of AI as mere disposable property – is unacceptable. The model proposed in this study will help to give rise to new ideas for governing this area, shifting the paradigm from mere post-hoc human liability to proactive systemic stewardship and nuanced responsibility allocation.

AI is such a complex subject that the established tendency of classifying every program as so-called "weak AI" is inappropriate. Currently, legislators in many countries are merely discussing one-sided protection thus guarding the rights of the program's copyright holder.[3] However, looking a little further, one can face a ques-

[3] *Artificial Intelligence Act* - 2021/0106(COD) (March 13 2024). Legislative Observatory of the European Parliament. https://oeil.secure.europarl.europa.eu/oeil/popups/summary.do?

tion of: "How could this area be regulated once "strong AI" (an autonomous machine capable of performing more than any modern neural network) appears in the world?"

Almost all of the government strategies, state plans, laws, regulations, decrees and other regulatory legal acts (hereinafter referred to as RLA) always pursue the same overriding goal – "to protect human rights".[4] Current research will not deviate from this idea, as its author believes that human rights and life always take precedence. This raises another question: "What is the reason for seeking ways to protect AI if it is not human and will most likely receive neither legal personality nor citizenship?" The point is that by failing to properly protect AI rights, humans endanger themselves. After all, there is always a risk of data corruption in this digitalized world.

REVIEW OF SOURCES

Basic terminology

This chapter should begin with a brief introduction to terminology. Without an understanding of the concepts outlined below, navigating the analysis of theoretical and empirical data will be quite challenging.

1) Artificial intelligence (also known as "AI" or less commonly "AIS" – artificial intelligence system) is a set of technological approaches that make it possible to simulate human cognitive functions, including finding solutions without using a predetermined algorithm, and obtaining results comparable or superior to those achieved as a result of human intellectual activity in performing specific tasks.

2) A dataset is data required for the development of computer programs based on artificial intelligence (AI), structured according to specific criteria and complying with the requirements of Russian legislation.

3) Strong artificial intelligence is a type of artificial intelligence system that can independently perform tasks and interact with people without human intervention.

---

id=1780523&t=e&l=en, retrieved March 8, 2026

[4] Agency's Artificial Intelligence (AI) Plan (May 14, 2021). The United States Agency for International Development. Agency Artificial Intelligence (AI) Plan | Open Government | US Agency for International Development (usaid.gov), retrieved March 8, 2026

Such systems are theoretically capable of adapting to changing circumstances.[5]

Literature review

In recent years, Russia has faced a significant surge in scholarly and regulatory attention to artificial intelligence.[6] However, most of the governmental efforts are focused on scientific researches rather than on developing a framework for legislative support for the use of AI tech. Given that the issue of interpreting the results of artificial intelligence is not a "blank spot" in legislation, it falls within a fairly well-developed area of law. The Russian scientific environment does not possess so many articles on the topic of determining the protectability and preservation of AI rights. Nevertheless, the key ones were included in this list, and a brief analysis of the most important ideas was conducted.

Belyakov, Khvorostyanaya, and Novikova (2021) depart from the ideas of a specific solution to the issue of AI authorship pursued by most other Russian researchers in this field. Instead, they propose a model of regulating AI actions that allows preventing future conflicts by recording planned actions in a virtual registry.[7] This proposal aims to change the established practice, where the legal system does not adapt to changes in society, but rather promotes the emergence of new legal relations using predictive methods.

Mikhovich E.A. in his work[8] raises the question of how exactly AI should be assessed – whether it should be regarded as a legal entity or as an animal. In addressing this issue, the author draws on the theory of "strong" and "weak" artificial intelli-

---

5 "National Strategy", approved by Decree of the President of the Russian Federation of October 10, 2019 N 490 (as amended on February 15, 2024), Section I, Art. 5, pp. 5-8

6 Reuters Staff. (2023, September 22). Governments race to regulate AI tools. Reuters. https://www.reuters.com/technology/governments-race-regulate-ai-tools-2023-09-22/, retrieved March 8, 2026

7 K. O. Belyakov, A. S. Khvorostyanaya, Yu. A. Novikova. (October 30, 2020). *Questions of the boundaries of subjective rights and freedoms and the boundary conditions of applicability to humans, cyborgs, and robots*. Innovations. No. 10 (264)

8 E.A. Mikhovich. (May 2022). Legal regime of works created using artificial intelligence. Questions of student science. No. 5 (69)

gence.[9] Analyzing the text of the article, one can clearly conclude that the author himself is more confident that strong AI does not yet exist and is unlikely to appear in the nearest future. It is also important to note that both authors cite the same Russian Civil Code article in their arguments, namely, article 1228 ("The author of the result of intellectual activity")[10]. Nevertheless, researchers interpret this provision in completely different ways: the former of the authors evaluate the contribution a human makes to the work performed by a robot.[11] The second one, on the contrary, looks at the process of AI activity in this way: the robot does not create, but only carries out what it is told, just like what a computer program used by a person does.[12]

The doctrinal debate about the security of AI ultimately returns to a fundamental question that most legal systems are not yet able to answer: "If an AI system leads to harmful or culturally destructive results, who (or what) can legitimately be held accountable?" Russian science, as reflected in the various interpretations of article 1228 of the Civil Code, is still trapped in a binary concept that seeks to determine the quantity of the human contribution to machine performance. While being useful for resolving traditional copyright disputes, this approach does not take into account the structural reality that advanced artificial intelligence systems can operate with a degree of autonomy that makes a simple human definition of authorship insufficient.

2. ANALYSIS OF REGULATORY AND OTHER OFFICIAL ACTS OF THE RUSSIAN FEDERATION

2.1. Review of Russian legislation

This chapter should be started with a discussion of one of the most important regulatory legal acts on AI of the Russian Federation: the National Strategy for the

[9] Ibid., pp. 391-392

[10] The Civil Code of the Russian Federation (Part one): No. 51-FZ dated 11.30.1994 (as amended on 12.16.2019). Collection of Legislation of the Russian Federation. May 12, 1994. – No. 32. – St. 3301

[11] Belyakov K.O., Khvorostyanaya A.S., Novikova Yu.A. Op. cit. P.2

[12] Mikhovich E.A. Decree. Op. P.392

Development of Artificial Intelligence for the Period up to 2030 (Decree of the President of the Russian Federation of 10.10.2019 No. 490 (as amended on 15.02.2024) "On the Development of Artificial Intelligence in the Russian Federation" (together with the "National Strategy for the Development of Artificial Intelligence for the Period up to 2030")) (hereinafter referred to as the "Strategy")

As stated in the general provisions of the document:

> This Strategy defines the goals and main objectives of the development of artificial intelligence in the Russian Federation, as well as measures aimed at its use in order to ensure national interests and implement strategic national priorities, including in the field of scientific and technological development.[13]

The goal of the strategy is set to:

> *Create a comprehensive system of legal regulation of public relations connected with the development and the use of AI technologies so also to ensure the safety of using such technologies.*[14]

However, this document does not address how exactly AI (which benefits are being described in every minute detail by the Legislator) should be protected.

The Strategy contains one significant section entitled as "Stimulating the Implementation of Artificial Intelligence Technologies in Economic and Social Sectors". This chapter examines the integration of AI systems into various areas of government activity. That includes: national projects, strategic areas of digital transformation in economic sectors, public administration, state-owned corporations, etc. In this regard, the Strategy completely contradicts the provisions set forth in the document on which it is based – the "Code of Ethics in the Sphere of AI". This section clearly reflects Governmental vision for the legal and regulatory framework for artificial intelligence activities. However, this provision focuses exclusively on one-sided protection of human rights and data, yet neither on who protects the AI itself, nor on who prosecutes the perpetrators. It should be as well noted that this regulatory legal

---

[13] Decree of the President of the Russian Federation dated October 10, 2019 N 490, section II, article 1, p. 4

[14] Decree of the President of the Russian Federation of 10.10.2019 N 490 (as amended on 15.02.2024), Section V, Article 24, paragraph G, Page 17

act is nothing more than a plan that should be implemented by year 2030. Consequently, the ideas outlined above cannot be considered absolute norms.

However, the development-focused idea of the National Strategy, although consistent from a strategic standpoint, exhibits regulatory inconsistency that manifests itself in the light of new risk-based approaches to regulation of artificial intelligence in foreign countries. With AI being defined mainly as a tool for accelerating economic modernization and strengthening the state apparatus, the strategy implies viewing artificial intelligence as an inert technology rather than an active element of socio-legal ecosystems. This innovation-focused approach, popular among developing digital economies, pays more attention to speed of innovation rather than institutional readiness for potential accountability issues. According to research literature on innovative economies, a lack of liability frameworks leads to negative regulatory externalities related to the integration of artificial intelligence technology, especially if it results in production of culturally valuable products or utilization of public infrastructures by algorithms. In the Russian context, the current lack of legislation regulating modifications, deletions, or errors related to AI should be viewed not as an omission but as a systemic characteristic of an innovative approach that relies on the assumption that human oversight will automatically follow AI implementation. However, increasing autonomy of neural networks and unexplainability of their decision-making makes this assumption legally unsustainable.

2.2. A review of official texts on AI development in the Russian Federation that are not of a regulatory nature

The "Code of Ethics in the Sphere of Artificial Intelligence" is a document that currently has no analogues in Russia. The aforementioned "Strategy" is based on it.[15] Unlike the previously cited Presidential Decree No. 490 and Government Order No. 2129-r, this document is only advisory in nature and does not oblige any person who has not "joined" this Code to comply with the instructions specified therein.

Considering the main provisions of this Code, which may correlate with the

[15] Code of Ethics in the Field of Artificial Intelligence (2021). Alliance in the Field of Artificial Intelligence. Code of Ethics 20101.pdf (a-ai.ru), retrieved March 8, 2026

problems identified earlier, the document states: "A person (an individual or legal entity recognized as a subject of liability in accordance with the current legislation of the Russian Federation) must always be held responsible for all consequences of the work of the AIS".[16] This provision can be considered essential for the protection of AI, since it demonstrates the true legal "attitude" of legislators towards artificial intelligence: AI is always right (a presumption of systemic reliability). However, even this principle has a number of issues:

Firstly, it is worth paying attention to the fact that the following Code "applies only to civil developments"[17] (as stated on the official website dedicated to the principles of this code). It can be safely concluded that these provisions will not apply to any state or pro-state developments of AI systems, neural network programs or other computers with integrated artificial intelligence. Therefore, returning to the problem discussed at the very beginning of the article: the problem of assigning blame not to humans but to machines and thus the risk of such an action.

The second problematic issue is status of the Code – "advisory": in order for the Code to enter into force, it is necessary for an entity or an individual to "join" it.[18] This means that the provisions contained therein cannot be mandatory. Furthermore, a perfectly reasonable question arises: "What (which legal act) regulates relations with AI for persons who have not signed into this Code?"

The Russian legal and regulatory framework, grounded in the National Strategy for Artificial Intelligence Development until 2030 and the advisory Code of Ethics in the Sphere of AI, reflects a global trend of emphasizing human-centric safeguards while sidestepping the issue of defining the legal status of AI. The strategy highlights economic integration and technological advancement but fails to address mechanisms for allocating responsibility when AI systems cause harm or their outputs are manipulated, altered, or deleted. The Ethics Code, by asserting that humans must always bear responsibility for all consequences of AI operations, effectively reinforces the problem it aims to solve: by insisting on human accountability in all instances, it perpetuates a legal fiction that grows increasingly untenable as AI systems become

[16] Ibid. Art. 3.2, p. 6

[17] Ibid. P. 1

[18] Ibid., Art. 2, p. 9

more autonomous and less reliant on explicit human instructions. Russia's current approach offers no comparable middle ground, leaving a regulatory void that could undermine both innovation and public trust. To address this, Russian law must move beyond the binary of "human blame versus AI immunity" and develop a tiered framework that recognizes varying degrees of AI autonomy and assigns corresponding obligations to developers, operators, and oversight bodies.

3. ANALYSIS OF INTERNATIONAL AND FOREIGN REGULATORY AND LEGAL ACTS ON AI

3.1. The approaches proposed in European Union

On 13 March 2024, the European Parliament adopted a resolution concerning a new Regulation of the European Parliament and the Council establishing previously agreed rules in the field of AI[19] (Artificial Intelligence Act, proposed by the European Commission back in 2021).[20] This event is being called the approval of the world's first AI Law. The aim of this Act is to improve the functioning of the internal market and promote the implementation of human-centered and trustworthy AI, while ensuring a high level of protection of health, safety, fundamental rights, innovations, environment and the rule of law. The following systems have been added to the list of high-risk systems for use:[21]

- as security components in management and operation of digital infrastructure, road traffic and the supply of water, gas, heating and electricity;[22]
- to determine access, enrollment or distribution of individuals in educational and vocational institutions at all levels;[23]

---

[19] Artificial Intelligence Act of January 24, 2024. The Council of the European Union. https://data.consilium.europa.eu/doc/document/ST-5662-2024-INIT/en/pdf, retrieved March 8, 2026

[20] Artificial Intelligence Act - 2021/0106(COD) (March 13 2024). Legislative Observatory of the European Parliament. https://oeil.secure.europarl.europa.eu/oeil/popups/summary.do?id=1780523&t=e&l=en, retrieved March 8, 2026

[21] Ibid., pp. 11-93

[22] Ibid., p. 34, p. 38

[23] Ibid., p. 32, pp. 36-37

- for the recruitment or selection of individuals or for the publication of targeted job offers, analysis and filtering of applications and evaluation of candidates;[24]
- to assess the eligibility of individuals to receive basic social security benefits and services, including health care;[25]
- to assess risks and determine prices for life and health insurance for individuals;[26]
- in the context of migration, asylum and border management, for the purposes of identifying or recognizing natural persons;[27]
- to influence the outcome of an election or referendum as well as the voting behavior of individuals when they exercise their right to vote[28].

Such systems must assess and mitigate risks, maintain usage logs, be transparent and accurate, and ensure human oversight. EU citizens will have the right to lodge complaints against AI systems and receive clarification regarding decisions made by high-risk AI systems that affect their rights. Unlike the "Code of Ethics" developed by Russian experts, virtually all provisions of EU AI Law are aimed solely at protecting human interests, and not a single word in the text says how exactly AI systems should be protected. It's worth noting that the EU has already considered the possibility of granting artificial intelligence legal personality. In 2018, the European Parliament rejected the idea of granting artificial intelligence legal personality.[29]

The EU Artificial Intelligence Act implements the principle of accountability through a risk-based mechanism that does not grant machine legal personality but instead places compliance obligations on developers and operators based on risk levels. The act eschews the metaphysical discussion of AI autonomy and instead

---

[24] Ibid., p. 40(a), p. 43

[25] Ibid., p. 16, pp. 25-26

[26] "Law on Artificial Intelligence" of January 24, 2024, paragraph 58(g), p. 57

[27] Ibid., p. 39, pp. 41-42

[28] Ibid., p. 40(a), p. 43

[29] The EU has the right to deny legal personality to artificial intelligence // European media platform "Euractiv" [electronic resource]. https://www.euractiv.com/section/digital/opinion/the-eu-is-right-to-refuse-legal-personality-for-artificial-intelligence/, retrieved May 5, 2026

introduces mechanisms such as transparency logging, human involvement, and post-market surveillance for high-risk AI. Instead of grappling with philosophical discussions regarding the ontology of AI systems, the act recognizes the impracticality of attributing blame to autonomous machines and focuses its regulations on individuals who have some form of responsibility in their creation and deployment. While comparing the EU's risk-based act with Russia's innovation-focused Strategy, one can observe that both jurisdictions face the same dilemma of regulating autonomy without providing it with any rights. In other words, the two approaches illustrate how comparable jurisdictions approach the same issue by diverting attention away from the debate surrounding liability towards human stewardship roles.

3.2. Approaches proposed in the USA

As stated earlier, the EU Law on Artificial Intelligence can be considered a breakthrough in this area, compared to the legislative frameworks of other countries and particular member states. Nevertheless, even here, the situation is not so clear-cut. It is worth saying that at the hearings in the US Congress[30] in 2023 Gary Marcus (AI expert) and Sam Altman (CEO of OpenAI – a company engaged in leading AI developments) both called for the US government to create new international governing bodies specifically for artificial intelligence similar to those regulating nuclear safety (the International Atomic Energy Agency – IAEA).[31] This only confirms the global governments' recognition of the importance of artificial intelligence technologies in the lives of future generations. The first try of creating an official legal framework for the development of AI to apply through the United States was the so-called "AI Bill of Rights", which was published in October 2022. It is a result of a collaboration between Office of Science and Technology Policy (OSTP), academic scientists, advocacy groups, general public, and even major companies like Microsoft and Google. It proposes ways to make AI more transparent, less discriminatory, and safer to use. It

[30] Kang, C. (2023, May 16). OpenAI's Sam Altman Urges A.I. Regulation in Senate Hearing. The New York Times. https://www.nytimes.com/2023/05/16/technology/openai-altman-artificial-intelligence-regulation.html, retrieved March 8, 2026

[31] Official website of the International Atomic Energy Agency [electronic resource]. Access mode: https://www.iaea.org/, retrieved May 5, 2026

also addresses current and potential harm to civil rights from AI, particularly in areas such as employment, education, healthcare, access to financial services and commercial surveillance. The Bill includes a set of five general principles[32], which can help to guide the development and the deployment of AI systems. All of it is designed with the civil rights of American society in mind. Each principle contains examples and practical steps that companies, governments, and other organizations can follow to incorporate these safeguards into their own policies and practices, as well as recommendations for overall technology development. Here is a brief look at each of them:

- Safe and efficient systems
- Algorithmic discrimination
- Data privacy
- Notification and explanations
- Human alternatives, consideration and fallback

As stated earlier, the Bill can be considered the first attempt of unifying legal regulation of AI. Before it, the US only had a multitude of unique AI laws passed at the local level.

Artificial intelligence laws passed in different states:

If it is being scaled down to the level of states, one can see that the approach is a little different with laws slowly but inevitably being created to address specific issues related to AI systems. For example, the state of Colorado passed a law,[33] which regulates how insurers can use big data and predictive models based on artificial intelligence to protect consumers from unfair discrimination. California, however, prohibits a person or an entity from using a chatbot to sell something to anyone or to influence their vote without disclosing the fact that it is in fact a chatbot.[34] Illinois became the first state to impose restrictions on the use of AI in hiring. Meanwhile, in 2021 New York City passed a law that requires companies to notify job seekers that they

---

32 Artificial Intelligence Bill of Rights Outline - October 2022, pp. 5-7

33 Limiting Insurers' Use of External Consumer Data. The Colorado State Assembly. Restrict Insurers' Use Of External Consumer Data | Colorado General Assembly, retrieved March 8, 2026

34 Senate Bill 1001. State Government of California. Bill Text - SB-1001 Bots: disclosure. (ca.gov), retrieved May 5, 2026

use a hiring algorithm. These corporations must have some independent auditors who annually review the technology for bias.[35] Authorities of a major number of states, including the ones listed, have attempted to ban or restrict the use of facial recognition technology by police.

Around 25 more states, including Puerto Rico and Washington, D.C., have introduced AI-related bills in 2023. States like Connecticut and Texas have even created policy bodies to monitor local AI developments. Some experts believe pressure on the federal government will likely increase as the US take AI regulation into their own hands.

3.3. Approaches proposed by UNESCO[36]

Now it shall be considered the ideas and provisions proposed by UNESCO (the main specialized agency of the United Nations for education, science and culture) in the context of the issue of AI set out in a document entitled "Recommendation concerning the Ethical Aspects of Artificial Intelligence".[37] One of the key objectives of the Recommendation is: *"…to promote equitable access to advances and knowledge in the field of AI…"*.

However, none of the 141 points of this document addresses how exactly the task of promoting equitable access to advances and knowledge in the field of AI should be addressed. Meanwhile, it should not be forgotten that one of UNESCO's primary objectives is preserving the world's cultural heritage and promoting technological progress. The only approach outlined that reflects this is the principle of transparency in the processes involved in any given program, including its development, training, and operation.

The UNESCO Recommendation on Artificial Intelligence has carved out an

---

35 New York May Regulate Hiring Algorithms. Community for national startups and technology companies. NYC May Regulate Hiring Algorithms. Here's What That Means. | Built In NYC

36 UNESCO (United Nations Educational, Scientific and Cultural Organization)

37 Recommendation on the Ethical Aspects of Artificial Intelligence (2021). UNESCO Digital Library. https://unesdoc.unesco.org/ark:/48223/pf0000380455_rus/PDF/380455rus.pdf.multi, retrieved March 8, 2026

entirely different normative sphere, emphasizing cultural preservation and human dignity rather than mere technical conformity. Although the ethical principles it espouses have no binding effect, the underlying notion of advanced AI systems being a new medium through which the heritage is transmitted directly strengthens the assertion about the public nature of some information-related programs, as argued before. The change of perspective that UNESCO brings about through its Recommendation – the question is no longer "Who owns the output?" but rather "Who ensures the preservation of the system?" – points to another weakness of the current Russian approach, namely the lack of mechanisms ensuring the preservation of culturally valuable AI systems. It should be emphasized again that mere ethical guidelines are incapable of protecting the heritage from being deliberately deleted or altered in any other way. For the UNESCO approach to be integrated into Russian doctrine, the normative requirements should be transposed into legal obligations regarding, for example, the necessity of creating archival backup copies and monitoring system changes.

3.4. Approaches proposed in other countries

In this section the author would like to highlight the most significant positions on issues of resolving disputes about AI proposed by other countries, in addition to those already mentioned earlier.

For example, in Estonia, a proposal to adopt a special AI law aimed at granting AI legal personality was rejected, indicating a tendency to oppose this idea shared among European countries.

China introduced a series of temporary measures that came into operation in summer 2024, requiring service providers to conduct security assessments and obtain approvals before releasing AI products to the mass market.[38] It's also worth saying that, according to the Chinese government, generative AI regulations will only apply to products made for the public, e.g. generative chatbots and public-facing

[38] o) Ye, J. (2023, July 14). China says generative AI rules to apply only to products for the public. Reuters. https://www.reuters.com/technology/china-issues-temporary-rules-generative-ai-services-2023-07-13/, retrieved March 8, 2026

models.

Australia will force search engines to develop new codes to prevent the dissemination of child sexual abuse material generated by artificial intelligence.[39]

Recently, the South Korean government noted the importance of reforming new laws and institutions needed to ensure the security of information technology systems. Furthermore, they expressed a position on the need to revise current legislation in order to recognize the rights and responsibilities of so-called "*electronic persons*".[40]

A comparative examination of international regulatory initiatives reveals a striking consensus on one point: no major jurisdiction is currently prepared to grant AI systems independent legal personality. The EU's Artificial Intelligence Act, the US AI Bill of Rights, UNESCO's ethical recommendations, and emerging national frameworks in China, South Korea, and elsewhere all operate on the implicit assumption that responsibility must ultimately rest with human actors: developers, deployers or users. This shared idea creates a structural vulnerability: as AI systems become more complex and their decision-making processes less transparent, the task of identifying a culpable human actor becomes increasingly difficult, if not impossible. The Estonian refusal to endow AI with personality, the Chinese requirement for a pre-market security assessment, the Australian obligation to moderate potentially harmful content, and the South Korean organizational changes indicate the way in which the jurisdictional authorities develop the concept of surrogate accountability when regulating their relationship with AI. When considered in comparison with each other, these examples clearly demonstrate a tendency: jurisdictions do not try to hold AI accountable because they distribute liability throughout its existence process. Russia's regulatory posture aligns with this global trend, but its reliance on non-binding ethical guidelines and development-focused strategies leaves it particularly

---

[39] Kaye, Ron. "Australia to Require AI-Made Child Abuse Material Be Removed from Search Results." Reuters, 11 Sept. 2023, www.reuters.com/technology/australia-require-ai-made-child-abuse-material-be-removed-search-results-2023-09-08/, retrieved March 8, 2026

[40] 11. International National Strategies for AI Regulation. Observer Research Foundation. https://www.orfonline.org/wp-content/uploads/2018/11/Ai_Book.pdf, retrieved March 8, 2026 P. 25

exposed to the risks of accidental damage. The comparative lesson is that more nuanced mechanisms for distributing responsibility across the AI lifecycle should be developed. This might include mandatory audit trails, shared liability regimes for high-impact systems or even specialized adjudicative bodies capable of evaluating technical evidence. By learning from the gradual approaches taken elsewhere while tailoring solutions to Russia's legal tradition, policymakers can begin to close the accountability gap without resorting to concepts like "electronic personhood".

4. REVIEW OF JUDICIAL PRACTICE

4.1. Analysis of international practice in resolving AI disputes

In March 2024, an American publisher "New York Times" covered an unusual incident from the world of law, in which artificial intelligence played a key role.[41] This case can be considered an example of abuse of AI and the shifting of responsibility to a robot:

A Manhattan district judge has refused to impose sanctions on Michael D. Cohen after he mistakenly provided his lawyer with fake legal precedents.[42] The submitted documents were generated by Google Bard, an AI program. They were intended to be included in a court filing. In December 2023, Mr. Cohen testified under oath that he had no idea that he "did not realize that Google Bard was a generative text service that, like ChatGPT, could show citations and descriptions that looked real but actually were not." Mr. Cohen thus avoided prosecution for violating trial rules and lying in court, citing his lack of knowledge of how the neural network worked.

Analyzing the judicial practice of China, Great Britain, and the United States, V.O. Kalyatin, author of the report "The Problem of Machine Creativity in the Legal System", concludes that, nowadays, courts around the world are still feeling negative-

---

41 Weiser, B. (2024, March 20). Judge Won't Punish Michael Cohen for Relying on Artificial Intelligence. The New York Times. https://www.nytimes.com/2024/03/20/nyregion/michael-cohen-artificial-intelligence.html, retrieved March 8, 2026

42 (March 20, 2024). United States of America v. Michael Cohen. The United States District Court for the Southern District of New York. https://storage.courtlistener.com/recap/gov.uscourts.nysd.499665/gov.uscourts.nysd.499665.108.0.pdf, retrieved March 8, 2026

ly about the idea of recognizing the AI rights, placing an emphasis on finding human contribution instead.[43] While society is only at the dawn of understanding the importance of AI programs intellectual property rights protection, the author has proposed several ideas for addressing this issue. These include a legislative distinction between intellectual property rights created by artificial intelligence itself and those created with its participation, recognition of intellectual property rights for the person managing the use of AI and the introduction of a special exception for the artificial intelligence usage for technological development and education.

The author confirms the tendency of increased interest in the development of artificial intelligence and machine creativity at international level. However, there is currently no unified document on the legal regulation of AI. The search for optimal legal frameworks is underway at three levels: international, supranational, and international-private[44]:

1) International level – involves consideration of the state's initiatives within the framework of UNESCO, WIPO[45], OECD[46], etc.;

2) Supranational – countries associations based on various principles, e.g., economic or political unions;

3) International-private – a level that unites entire institutions and private technology companies uniting institutions; e.g., the European Alliance for Research Excellence (supported by Microsoft) and the International Federation of Robotics (IFR)

4.2. Analysis of trends in AI dispute resolution

At the beginning of this paper, the author has already presented some arguments for why the current AI legal status disputes resolving "trend" may not be sufficient to address the vast number of cases that may arise in the future. This trend can be simply formulated in the following way: courts are not focusing on what the EU and

43 V.O. Kalyatin, p. 45, p. 18

44 Kalyatin V.O. Decree. cit., paragraph 27, pp. 10-11

45 World Intellectual Property Organization (hereinafter referred to as WIPO)

46 Organisation for Economic Co-operation and Development (hereinafter referred to as OECD)

UNESCO stipulate in their legal acts, but cling to the strategy where they keep identifying the human contribution in any result of the intellectual activity of robots and AI in general. This is a "primitive" position, which cannot be reflected in the future, when a proposed technology known as "strong AI" would appear in modern world. However, officials all around the world are already working on the idea of introducing new approaches. For example, the Russian government has developed a mechanism to protect citizens from harm caused by artificial intelligence technologies.[47]

According to the proposed amendments to the Law "On Experimental Legal Regimes in the Sphere of Digital Innovations", a specially organized "commission" will be responsible for investigating the case in which abuse of AI occurs or the harm was caused through the use of artificial intelligence.[48] For the model's final development, some of the presented ideas could be borrowed: e.g., the proposal to assert the absolute AIS innocence for any illegal actions committed using the results of the program's intellectual activity.

Maksim Milkov, the head of the artificial intelligence practice at Softline Digital, believes that additional layers of protection should be used to improve the experimental legal regime. For example, in cases where the AI has made a "mistake", he says, a separate issue concerns liability for the consequences caused by AI malfunctions. However, the lawyer himself cannot answer the question of "Who exactly should bear responsibility in such a case: the developer or the program itself?" In addition to the expert's thoughts, one could begin to reflect on another pitfall of error in the field of AI regulation – the gradation of such an error. Systems operate with probabilities and the risk of error is always acceptable. Nevertheless, it's not always possible to clearly identify a genuine error and distinguish it from a so-called hack (information leak or deliberate distortion of data included in the AIS dataset).

Contemporary judicial practice across multiple law systems demonstrates a

---

[47] m) Tkachev, I. (2023, December 7). Authorities have developed a mechanism to protect against harm caused by AI technologies. RBC News Agency. https://www.rbc.ru/economics/07/12/2023/657063269a79470fdb112ac3, retrieved March 8, 2026

[48] Federal Law of July 31, 2020 No. 258-FZ On experimental legal regimes in the field of digital innovation in the Russian Federation. Internet portal of legal information. http://pravo.gov.ru, retrieved March 8, 2026

conservative reliance on human-centric liability frameworks. This tendency risks creating a perverse incentive structure: if harm caused by AI can always be deflected onto an unwitting human actor, there is little pressure on developers to build more transparent, accountable or resilient systems. Russia's proposed experimental legal regimes and harm-compensation commissions represent a promising step toward addressing this gap. To align with emerging global best practices, Russian jurisprudence must evolve from post-hoc human attribution toward proactive institutional oversight.

5. POSSIBLE APPROACHES TO PROTECTING AI RIGHTS

5.1. Foundations for the development of a special model for regulating the protectability of AI in the Russian Federation

In order to fully understand the grounds for author's desire to propose changes to some AI-related regulation provisions preserved in Russian law, it's worth examining another AI technology.

Humanity's eternal dream of increasing data storage capacity while reducing its physical parameters is becoming increasingly realistic. These days, many AI developments are focused on using neural networks and artificial intelligence to store digital information. To make such ideas come true, digital scientists are proposing the use of ML (Machine Learning) technology, which allows robots to be trained not directly but through the use of similar problem-solving techniques.[49] One of the representatives of such breakthrough technologies in the field of data storage is the "AiRE" system. It is a unified data platform that provides efficient storage and management of data across various memory locations and storages, from local disks to cloud and peripheral devices.[50]

Experts in the field of AI technologies claim that it is indeed possible to carry out the process of "*Forcing AI to 'forget' some information*".[51] The humanity is dealing

[49] 7 Ways Artificial Intelligence Will Transform Data Storage. Dataversity. 7 Ways AI Will Transform Data Storage - DATAVERSITY, retrieved May 5, 2026

[50] AiRE platform website. https://storvix.eu/, retrieved May 5, 2026

[51] Snyder, A. (2024, January 12). Machine forgetting: How difficult it is to get AI to forget. Axios.

with the risk of data distortion, which in some cases can lead to the corruption of history and substitution of cultural values. It should still not be forgotten that defining responsibility of state-segment used AIS is impossible without conducting a full transparent analysis before accusing the system of making a mistake that led (or could potentially lead) to some negative actions towards citizens or the state. Here is an example: a lot of scientists claim that just in a few centuries a true masterpiece of fine arts and One of Leonardo da Vinci's most famous paintings, The Mona Lisa is going to completely vanish due to pigment evaporation and chemical varnish deformation, which makes the paint fade away slowly. In the nearest future, some software developer can possibly develop a new AI-powered program, working as a vast and complicated storage for digitized information. Thus, he can upload the great portrait into his creation. In the real world there will no longer be any trace of the UNESCO guarded object. The only "authentic" copy will be stored in this program. Afterwards, if the developer decides to exercise his right to delete his program, the question asked earlier will arise again: "Does he actually have the right to commit this action?" Speaking from the point of view of modern Russian and foreign legislation, the answer is simple: "Yes, he does". After all, no legal act states otherwise. However, such actions are not fair, at least not to humanity, as all other people will be deprived of the opportunity to use this resource.

Second reason for removing an AI program is public opinion. One can imagine a program that creates images in the style of early Impressionist artists based on user requests. Some customers might dislike the neural network's performance – as they find it wrong that the system imitates other artists, as it "devalues art in general". On the other hand, the Union of Artists might recognize this program as a very useful one for teaching the visual arts to younger generations. Therefore, it can be concluded that public opinion is not always sufficient grounds for negative actions against AI.

In addition, it is impossible to exclude the possibility of "hacking" AI, when a program with valuable information or benefits for society is used in a bad way. In the hands of malicious actors, a tool that was used for training yesterday could now

https://www.axios.com/2024/01/12/ai-forget-unlearn-data-privacy, retrieved March 8, 2026

become a weapon for coordinating strategic military attacks.

Touching on the topic of weapons, it's also important to note that currently there is no law that rules out the use of AI in military affairs. On the other hand, the norms outlined in the National Strategy to 2030 and the Code of Ethics in AI are based on A. Asimov's rules of robotics, which essentially indicate to a robot that it has no right to harm a human. Therefore, legislator can strive to minimize the risks of hacking into AI systems and using them in negative purposes by implementing the ethical Code in all areas of life where AI is used, including government research.

Furthermore, a more detailed approach to developing specific AI regulations regarding its protectability is proposed. This is necessary for two reasons:

Hacking an AI and using its output doesn't automatically render the program "unusable" – its utility and capabilities remain intact. So, the question is: "Should AIS be removed if its creativity is being misused?" And that is precisely the question that courts should be asking when considering whether to allow the program's removal.

If there is a need to integrate AI into public administration, then the protection of this very AI must be strengthened. After all, the current situation is such that no one who violates a program's rights is held accountable in court. Moreover, the program is subject only to enhanced restrictive measures, not to methods of protection against malicious actors.

Let's return to the main question of the study and consider it from a different perspective: "What if the issue is not deletion but modification of the program content?" In many cases, fraudulent substitution can be even worse than erasure. "Who is responsible for overseeing such modifications in the AI program?" Currently, all laws and official texts on this topic state that only a limited number of individuals, such as developers and operators, are allowed to work with the code of these programs. However, this does not contradict the information distortion risk for future generations. This risk arises from the fact that by using AI as a storage and having full access to the ability to modify the data kept within it, the owner (or another authorized individual) may distort the data. Such manipulations can have a negative impact on various aspects, including art, legal regulations, and even historical documents. This can also have a significant impact on politics, as the distortion of historical data can lead to information warfare. To address these concerns, the author proposes the

following approach:

If a crime is committed against an object protected by UNESCO and stored in an AI program, the perpetrator should be adjudicated before competent national courts with extraterritorial jurisdiction provisions, or referred to specialized international arbitration mechanisms analogous to those governing cultural heritage protection

If such an act is committed against a program used in government administration, the individual who violated the law must be brought before one of the Federal Courts of the Russian Federation, depending on the regulations proposed by the State;

CONCLUSION

To summarize, the article was able to solve the following problems:

Roughly outlined the problematic AI regulation aspects;

Studied and analyzed a set of cases, related to the legal artificial intelligence status;

Have put forward proposals on the basis of an approach to the formation of amendments and additions to the current legislation.

The works of contemporary experts were reviewed and their opinions were explored on the issues raised. Some studies and reports on the legal nature of artificial intelligence were revised. The legislation of the Russian Federation in areas related to the AI protection was then analyzed. In addition, approaches and recommendations from other countries, international unions and human rights organizations were studied. Then, various cases from foreign judicial practice were considered (as well as other precedents directly related to the resolution of controversial AI issues).

A detailed analysis revealed a certain tendency among courts, from which the judicial system, in the author's opinion, should be separated. Ultimately, the current approach to resolving the issue of AI rights cannot be used in the future, as it would not realistically reflect the ongoing situation. It should also be added that talking about AI's responsibility, which legislators are not fully excluding from the concept of AI regulation, can hardly be recognized as fully appropriate, given that there has been no talk yet about the rights of AI.

Taking into account the above said, proposals were put forward to modernize and supplement existing laws that would reflect all the approaches considered during the

study. Naturally, the provisions from this study cannot be amended or supplemented by existing legislation overnight. However, the proposals and developments discussed may be found extremely useful for future legislators and may provide some food for thought on the topic of AI. After all, one of the main objectives of this work was to demonstrate that the legal system must never fall behind in the rapidly evolving world of technology. The proposed approaches should in no way violate the norms previously adopted and discussed by legislative bodies because their goal is to serve merely as a useful supplement.

By facing advanced AI systems not merely as disposable tools subject to absolute property rights, but as unique digital assets, the law can establish a specialized regulatory approach. This framing ensures that the proposed obligations function as a legitimate, proportionate limitation on absolute ownership.

This study should be concluded with a quote from the great American scientist, theoretical physicist, and astrophysicist Stephen Hawking:

The creation of artificial intelligence may be humanity's last technological achievement if we don't learn to manage the risks.

This phrase perfectly conveys the importance of a responsible approach to artificial intelligence technologies. AI can be incredibly useful, but for this technology to work "for us", the humanity needs to learn to manage it, not only tightening the restrictions but also involving mutual protection of rights.

REFERENCES

Regulatory legal acts and other documents (not having regulatory legal status):

The Civil Code of the Russian Federation (Part one): No. 51-FZ dated 11.30.1994 (as amended on 12.16.2019). Collection of Legislation of the Russian Federation. May 12, 1994. – No. 32. – St. 3301

Federal Law of July 31, 2020 No. 258-FZ *On experimental legal regimes in the field of digital innovation in the Russian Federation.* Internet portal of legal information http://pravo.gov.ru, retrieved March 8, 2026

Decree of the President of the Russian Federation of October 10, 2019 N 490 (as amended on February 15, 2024) *On the Development of Artificial Intelligence in the Russian Federation (together with the National Strategy for the Development of Artificial Intelligence for the Period up to 2030)*. Collection of Legislation of the Russian Federation. October 14, 2019. N 41, Art. 5700

(August 19, 2020) Order of the Government of the Russian Federation dated No. 2129-r *Concept for*

the Development of Regulation of Relations in the Sphere of Artificial Intelligence and Robotics Technologies until 2024 Internet Portal of Legal Information. http://pravo.gov.ru, retrieved March 8, 2026

Recommendation on the Ethical Aspects of Artificial Intelligence (2021). UNESCO Digital Library. https://unesdoc.unesco.org/ark:/48223/pf0000380455_rus/PDF/380455rus.pdf.multi, retrieved March 8, 2026

Code of Ethics in the Field of Artificial Intelligence (2021). Alliance in the Field of Artificial Intelligence. Code of Ethics 20101.pdf (a-ai.ru), retrieved March 8, 2026

Artificial Intelligence Act - 2021/0106(COD) (March 13 2024). Legislative Observatory of the European Parliament. https://oeil.secure.europarl.europa.eu/oeil/popups/summary.do?id=1780523&t=e&l=en, retrieved March 8, 2026

Artificial Intelligence Act of January 24, 2024. The Council of the European Union. https://data.consilium.europa.eu/doc/document/ST-5662-2024-INIT/en/pdf, retrieved March 8, 2026

Artificial Intelligence Bill of Rights Plan (October 2022). The US Government. https://www.whitehouse.gov/wp-content/uploads/2022/10/Blueprint-for-an-AI-Bill-of-Rights.pdf, retrieved March 8, 2026

Explanatory Notes on the AI Bill Plan. The US Government. https://www.whitehouse.gov/ostp/ai-bill-of-rights/, retrieved March 8, 2026

International National Strategies for AI Regulation. Observer Research Foundation. https://www.orfonline.org/wp-content/uploads/2018/11/Ai_Book.pdf, retrieved March 8, 2026

Ethical Principles of Artificial Intelligence. US Department of Defense, Office of the Chief of the Digital Technologies and Artificial Intelligence Division. Microsoft Word - Ethical Principles for Artificial Intelligence (ai.mil), retrieved March 8, 2026

Agency's Artificial Intelligence (AI) Plan (May 14, 2021). The United States Agency for International Development. Agency Artificial Intelligence (AI) Plan | Open Government | US Agency for International Development (usaid.gov), retrieved March 8, 2026

Effective Machine Learning Practices for Medical Device Development: Guiding Principles (October 2021) The U.S. Department of Health and Human Services. Guiding Principles - GMLP (fda.gov), retrieved March 8, 2026

Federal Protections Against National Origin Discrimination (October 2000) The U.S. Department of Justice's Civil Rights Division. Civil Rights Division | Federal Protections Against National Origin Discrimination (justice.gov), retrieved March 8, 2026

The Federal Trade Commission's (FTC) Plan to Regulate AI. The US Federal Trade Commission on International Development. Aiming for truth, fairness, and equity in your company's use of AI | Federal Trade Commission (ftc.gov), retrieved March 8, 2026

Copyright and Artificial Intelligence. The US Copyright Office. Access mode: Copyright and Artificial

Intelligence | US Copyright Office, retrieved March 8, 2026

Limiting Insurers' Use of External Consumer Data. The Colorado State Assembly. Restrict Insurers' Use Of External Consumer Data | Colorado General Assembly, retrieved March 8, 2026

Executive Order of the President of the United States on Artificial Intelligence (October 30, 2023). The US Government. https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/, retrieved March 8, 2026

Judicial practice:

20. Zhou B. (2019). *Artificial Intelligence and Copyright Protection – Judicial Practice of Chinese Courts.* The World Intellectual Property Organization https://www.wipo.int/export/sites/www/about-ip/en/artificial_intelligence/conversation_ip_ai/pdf/ms_china_1_en.pdf, retrieved March 8, 2026

21. (March 20, 2024). *United States of America v. Michael Cohen*. The United States District Court for the Southern District of New York. https://storage.courtlistener.com/recap/gov.uscourts.nysd.499665/gov.uscourts.nysd.499665.108.0.pdf, retrieved March 8, 2026

Literature:

22. K. O. Belyakov, A. S. Khvorostyanaya, Yu. A. Novikova. (October 30, 2020). *Questions of the boundaries of subjective rights and freedoms and the boundary conditions of applicability to humans, cyborgs, and robots*. Innovations. No. 10 (264);

23. E.A. Mikhovich. (May 2022). *Legal regime of works created using artificial intelligence*. Questions of student science. No. 5 (69);

24. A.A. Beloshapkina. (2022). *Legal regulation of artificial intelligence in Russia and abroad*. Bulletin of the Magistracy. No. 10-3 (133);

25. V.O. Kalyatin. (2021). *The Problem of Machine Creativity in the Legal System: Regulation of the Creation and Use of the Results of Intellectual Activity Using Artificial Intelligence, Foreign Experience and Russian Prospects*. Publishing House of the Higher School of Economics;

Other sources (information from the media):

Sazonova M. (2020, July 16). *Artificial intelligence and law: is there a contact?*; Garant.Ru. https://www.garant.ru/news/1401154/?ysclid=lo6ssqr88o452055317, retrieved March 8, 2026

Kalincheva M. (25.02.2020). Zakon.ru. https://zakon.ru/blog/2020/02/25/o_nekotoryh_voprosah_pravovoj_ohrany_izobretenij_sozdannyh_iskusstvennym_intellektom?ysclid=lo6vicuxlx245502367, retrieved March 8, 2026

NYC May Regulate Hiring Algorithms. Here's What That Means. | Built In NYC. (2021). Built In. https://www.builtinnyc.com/articles/nyc-hiring-algorithm-bill-julia-stoyanovich?__cf_chl_rt_tk=_93brDQTW0Q1oG8M.zoqorfSgKrbYCvObFJyQRKOcNI-1772974571-1.0.1.1-ZVG8poktmZMrTVnGPpcbBvdKxzu8ytoVDI308kO0Caw, retrieved March 8, 2026

Perkova A. (2021, October 26). *Russian will be able to sue artificial intelligence*. URA.RU. https://ura.news/news/1052513302, retrieved March 8, 2026

Negi, S., & Scott-Briggs, A. (2024, October 18). *Dr. Wade Newman Grant Opens Doors for Aspiring Healthcare Professionals with $1,000 Award*. TechBullion. https://techbullion.com/the-art-exchangetaex-features-legendary-artist-manovich-in-the-latest-nft-drop/, retrieved March 8, 2026

Ramirez, J. (2023, September 5). *AI Mistakes: Who's Responsible? | Built In*. Builtin.com. https://builtin.com/artificial-intelligence/responsibility-for-AI-mistakes, retrieved March 8, 2026

Glover, E. (2024, March 19). *AI Bill of Rights: What You Should Know | Built In*. Builtin.com. https://builtin.com/artificial-intelligence/ai-bill-of-rights, retrieved March 8, 2026

Harris, D. E. (2023, June 16). AI is already causing unintended harm. What happens when it falls into the wrong hands? *The Guardian*. https://www.theguardian.com/commentisfree/2023/jun/16/ai-new-laws-powerful-open-source-tools-meta, retrieved March 8, 2026

Reuters Staff. (2023, September 22). Governments race to regulate AI tools. *Reuters*. https://www.reuters.com/technology/governments-race-regulate-ai-tools-2023-09-22/, retrieved March 8, 2026

Hamilton, S. (2024, February 21). *7 Ways AI Will Transform Data Storage - Dataversity*. Dataversity. https://www.dataversity.net/articles/7-ways-ai-will-transform-data-storage/, retrieved March 8, 2026

Kang, C. (2023, May 16). OpenAI's Sam Altman Urges A.I. Regulation in Senate Hearing. *The New York Times*. https://www.nytimes.com/2023/05/16/technology/openai-altman-artificial-intelligence-regulation.html, retrieved March 8, 2026

Weiser, B. (2024, March 20). Judge Won't Punish Michael Cohen for Relying on Artificial Intelligence. *The New York Times*. https://www.nytimes.com/2024/03/20/nyregion/michael-cohen-artificial-intelligence.html, retrieved March 8, 2026

Tkachev, I. (2023, December 7). *Authorities have developed a mechanism to protect against harm caused by AI technologies*. RBC News Agency. https://www.rbc.ru/economics/07/12/2023/657063269a79470fdb112ac3, retrieved March 8, 2026

Burri, T. (2018, May 31). *The EU is right to refuse legal personality for Artificial Intelligence*. Www.euractiv.com. https://www.euractiv.com/section/digital/opinion/the-eu-is-right-to-refuse-legal-personality-for-artificial-intelligence/, retrieved March 8, 2026

Ye, J. (2023, July 14). China says generative AI rules to apply only to products for the public. *Reuters*. https://www.reuters.com/technology/china-issues-temporary-rules-generative-ai-services-2023-07-13/, retrieved March 8, 2026

Kaye, Ron. "Australia to Require AI-Made Child Abuse Material Be Removed from Search Results." *Reuters*, 11 Sept. 2023, www.reuters.com/technology/australia-require-ai-made-child-abuse-material-be-removed-search-results-2023-09-08/, retrieved March 8, 2026;

Snyder, A. (2024, January 12). *Machine forgetting: How difficult it is to get AI to forget*. Axios.

https://www.axios.com/2024/01/12/ai-forget-unlearn-data-privacy, retrieved March 8, 2026

V
ARTIFICIAL INTELLIGENCE: BETWEEN FUNDAMENTAL RIGHTS AND SYSTEMIC RISKS

THE SECURITIZATION OF ARTIFICIAL INTELLIGENCE: NATIONAL SECURITY EXPANSION AND ITS INSTITUTIONAL IMPACT ON INVESTMENT SCREENING

Gulinigaer Wumiti

East China University of Political Science and Law

DOI: https://doi.org/10.59224/dti8.ch16

Abstract: This article examines how specific existential threats have shaped the U.S. national security concept, tracing its evolution across five historical transformations—from the Cold War to the current era of dual-use technology and geopolitical rivalry. The study focuses on how this expansion on national security concept directly shapes investment screening. Through a comparative analysis of the inbound CFIUS framework and the outbound OISP mechanism, the research demonstrates the weaponization of these regulatory tools. By specifically examining AI, the paper argues that the tightened scrutiny of AI investments is driven not only by the broader securitization trend, but also by the technology's core characteristics and external environment. Ultimately, this aggressive regulatory shift creates a chilling effect on investors, disrupts global tech governance, and slows the pace of innovation.

Keywords: National Security Expansion; Investment Screening; Artificial Intelligent; CFIUS; Outbound Investment Security Program.

Table of contents: INTRODUCTION; 1. The Five Transformations of the U.S. National Security Concept; 1.1. The First Transformation: The Cold War and the Focus on Ideological Threats; 1.2. The Second Transformation: The Nixon Era and the Balance of Power; 1.3. The Third Transformation: The Clinton Era and the Rise of Economic Security; 1.4. The Fourth Transformation: 9/11 and the Shift to Preemption; 1.5. The Fifth Transformation: Dual-Use Tech and the Weaponization of National Security; 2. Current Landscape of U.S. Investment Screening Mechanism; 2.1. Inbound Investment Screening: The CFIUS Framework; 2.2. Outbound Investment

Screening: The OISP Mechanism; 3. Analysis of the Drivers Behind the Tightening of AI Investment Scrutiny; 4. Assessment of the Tightened U.S. Investment Review System; CONCLUSION; References.

INTRODUCTION

Today, security threats are hard to see but exist everywhere. Large language models and other forms of artificial intelligence (AI) have significantly boosted efficiency in civilian industries. However, their military applications have simultaneously made AI a central focus of national strategy and security. Driven by this technological shift, the United States has continuously expanded its concept of national security, directly resulting in a tightening of investment regulations. This regulatory shift is clearly marked by recent policy actions, ranging from Executive Order 14083 in 2022 to the Treasury Department's Final Rule on outbound investment (31 CFR Part 850) issued in late 2024. This article examines how this conceptual expansion plays out in investment screening.

This article aims to trace how the concept of national security has evolved. By doing so, it demonstrates that the overexpansion and weaponization of this concept form the underlying logic behind today's strict investment rules. In other words, the tightening of both inbound and outbound investment screening is simply the institutional extension of a broadened security concept. Using artificial intelligence as a concrete case study, this article examines the drivers behind these stricter controls and analyzes their consequences—not only for individual investors, but also for the broader international order, geopolitical competition, and global AI governance. Ultimately, this article argues that as national security becomes weaponized, investment regulatory mechanisms are increasingly reduced to geopolitical tools. Rather than serving as a defensive shield for "protection," they now function as an offensive weapon for "attack."

Part I begins by analyzing how specific historical events and existential threats have triggered transformations in the U.S. national security concept. Following this overview, Part II examines the regulatory processes, rationales, and structural flaws of both inbound and outbound U.S. investment screening mechanisms. By con-

ducting a longitudinal review of their development and a horizontal comparison of their features, this section clarifies how these systems operate. Furthermore, through an analysis of Ralls Corp. v. CFIUS and the Nippon Steel case, Part II assesses the practical challenges of enforcing these regulations and the viability of legal remedies for investors. Part III then narrows the focus to AI as a specific target of this tightened scrutiny. It explains the drivers behind this regulatory trend, analyzing how AI's unique features, U.S. domestic policies, and the broader international environment contribute to stricter oversight. Finally, Part IV evaluates the consequences of these restrictive policies. It focuses specifically on the chilling effect this regulatory tightening has on investors and its disruptive impact on the international order, particularly regarding global AI governance.

1. THE FIVE TRANSFORMATIONS OF THE U.S. NATIONAL SECURITY CONCEPT

The following section examines five critical stages in the historical evolution of the U.S. national security concept. This article argues that the formulation of national security is never accidental; the concept of "security" only holds meaning when contrasted with a "threat." Therefore, rather than directly identifying the turning points, this section first analyzes major historical threats. By examining the policy shifts, official documents, and diplomatic strategies triggered by these events, it traces exactly how the concept of national security transformed.

1.1 The First Transformation: The Cold War and the Focus on Ideological Threats

The first transformation of the U.S. national security concept can be focused on the Cold War era, with the National Security Act of 1947[1] serving as the clear reflection of this shift. Following the end of the Second World War, the global power split into two camps, formally establishing the Cold War structure. Beyond the shifts in the international power dynamic, the emergence of nuclear weapons fundamentally reshaped perceptions of peace and war. The first decade after the Second World War

[1] National Security Act of 1947, Pub. L. No. 80-253, 61 Stat. 495.

is described by David A. Baldwin as "the most creative and exciting period in the entire history of security studies."[2] Given that a securitized outcome arises from a specific existential threat, the threats that drove the transformation of the U.S. national security concept during this period were the rise of a new ideological power and the creation of devastating nuclear weapons. First, World War II destroyed the traditional balance of power in Europe. At this time, the Soviet Union not only held an ideology completely opposed to the United States, but also possessed massive military and industrial capabilities to export its ideology globally. The United States no longer viewed the situation merely as a power struggle between two nations, but rather as an ideological contest filled with uncertainty. Second, the introduction of nuclear weapons completely redefined human understanding of the outcomes of war. Before this, no weapon was capable of causing such massive, devastating, and unselective harm. Moreover, the outbreak of a nuclear war would result in mutual destruction, leaving no victors. As a result, nuclear weapons are by nature unpreventable; it is impossible to form an improvised response as one might against conventional weapons. The fact that such weapons were controlled by an adversary driven by a new ideology further deepened U.S. anxiety.[3]

Having analyzed the existential threat, it is necessary to focus the securitized outcome of this threat: what specific transformations take place on the concept of national security? In 1943, Lippmann stated, "A nation is secure to the extent to which it is not in danger of having to sacrifice core values if it wishes to avoid war, and is able, if challenged, to maintain them by victory in such a war." [4]In other words, what a nation defends through war shows its core national security. It is evident that the U.S. initially held a reactive posture toward threats; security was defined simply as the absence of situations threatening core values, and reactions happened only when a threat actually appeared, reflecting a certain immediacy. Furthermore, U.S. isolationism prior to World War II fostered ongoing debate over whether the country

---

[2] David A. Baldwin, quoted in Barry Buzan & Lene Hansen, The Evolution of International Security Studies 67 (2009).

[3] Nat'l Sec. Council, NSC 68: United States Objectives and Programs for National Security (Apr. 14, 1950) (reflecting deep-seated American anxiety and apprehension regarding the Soviet Union's acquisition of nuclear weapons).

[4] Walter Lippmann, U.S. Foreign Policy: Shield of the Republic 9–10 (1943).

should entangle itself in European affairs and military alliances.[5] From the perspective of national security, isolationism argued that as long as domestic affairs within specific territories were managed well, the geographic advantage would protect the nation from foreign conflicts. In summary, the pre-Cold War concept of national security was characterized by geographical reliance and reactive immediacy. Post-Cold War, however, the presence of a new ideology and nuclear weapons pushed the U.S. national security concept in a new direction.

This transformation appeared in several ways. First, the object of national security became more targeted. Previously, danger was viewed merely as an immediate concept; the United States narrowed its defensive focus to the ideological level, establishing a more clear target. Second, the establishment of the national security system extended from wartime into peacetime. [6]Driven by the new ideology and new weapons, the United States realized that waiting until danger arrived to build a national security system would be too late. As a result, it began preparing defensive frameworks during peacetime. One of the duties of the National Security Council, established by the National Security Act of 1947, was to "assess and appraise the objectives, commitments, and risks of the United States in relation to our actual and potential military power, in the interest of national security; for the purpose of making recommendations to the President in connection therewith."[7] This indicates that the United States was no longer passively reacting, but actively identifying and addressing crises. While establishing such a peacetime security system appears perfectly logical today, it provoked significant debate at the time[8]. Senator Edward V. Robertson of Wyoming warned that Congress was creating a "Frankenstein monster"[9] and urged his

---

5 Bear F. Braumoeller, The Myth of American Isolationism, 6 Foreign Pol'y Analysis 349, (2010).

6 Nicholas Steneck, National Security Act of 1947, in Routledge Resources Online - The Cold War (Ruud van Dijk & Liliane Stadler eds., 2025) (explaining that to wage a global struggle against communism, the United States had to maintain a massive peacetime national security system)

7 National Security Act of 1947, 50 U.S.C. § 3021(b).

8 Robert D. Russell II, The National Security Debate and the Truman Administration's Policy Toward China, 1947-1950 (2000) (analyzing two different opinions on how to respond to the menace of communism among officials inside and outside the administration of Harry S. Truman).

9 Nicholas Steneck, National Security Act of 1947, in Routledge Resources Online - The Cold War

colleagues to "not let our fear of Communism blind us to the danger of military domination." Third, the United States transitioned from addressing specific, imminent threats to preparing for long-term, systemic risks. The dangers posed by nuclear weapons and new ideologies were not immediate, meaning the nation could no longer afford to wait until these dangers happened before reacting. As a result, the object of securitization shifted from specific threats to long-term risks.

1.2 The Second Transformation: The Nixon Era and the Balance of Power

Following the first shift, the second transformation of the U.S. national security concept centered around the Nixon era. During this period, first, the United States was deeply entangled in the Vietnam War. This conflict not only drained U.S. financial and human resources but also weakened its military power, causing it to lose control over core strategic areas on the global stage. Second, following the Cuban Missile Crisis, the Soviet Union accelerated its nuclear weapons development. With both nations' nuclear weapons largely keeping each other in check, the U.S. and the Soviet Union formally entered an era of nuclear balance. Third, the internal dynamics of the socialist nations shifted during this time, highlighted by the split between China and the Soviet Union, as conflicts arose between the two nations over national interests. Finally, unchecked spending from the Vietnam War brought the Bretton Woods system to the brink of collapse. At the same time, European allies lost confidence in the United States, dividing the western allies.

Based on this analysis, the existential threat that drove the shift in the national security concept during this period came from the pursuit of ideological unity to achieve absolute security for the nation and the international order—a pursuit that resulted in near economic collapse and deep entanglement in war. Therefore, realists like Nixon and Kissinger realized that it was unrealistic to continuously reject and interfere in political systems different from their own.[10] Instead, they believed that power among different nations should keep each other in check to achieve harmony

(Ruud van Dijk & Liliane Stadler eds., 2025) (quoting Sen. Edward V. Robertson).

[10] Richard M. Jennings, The Thrust of the Nixon Doctrine,(1972).

and security.[11] In a 1971 interview with TIME magazine, Nixon stated: "We must remember the only time in the history of the world that we have had any extended period of peace is when there has been balance of power... I think it will be a safer world and a better world if we have a strong, healthy United States, Europe, Soviet Union, China, Japan, each balancing the other, not playing one against the other, an even balance." [12]Henry Kissinger, Secretary of State under President Richard Nixon, believed that the core of "Realpolitik" was to deal with major powers in a practical manner, free from the limits of strict political ideologies or moral guidelines.[13] Nixon's visit to China during this period served as the best proof of this theory. Building on the analysis of the existential threat above, the securitized outcome can be summarized as follows: the concept of national security began to move away from its strong ideological tone, prioritizing national interests over ideological differences, and redefining security as a balance of power among major nations.

1.3 The Third Transformation: The Clinton Era and the Rise of Economic Security

During the Clinton era, we can observe the third transformation of the national security concept. In the National Security Strategy of Engagement and Enlargement, this period is described as "a period of great promise but also great uncertainty,"[14] and indeed it was. During this time, the Soviet Union collapsed, and the United States emerged from the Cold War as the only "superpower" in the world. This era also saw the rapid development of globalization; the Clinton administration promoted the enactment of the North American Free Trade Agreement and GATT, fostered the global trading system, and China officially became a member of the World Trade Organization. At the same time, the rise of the Internet started the information technology revolution. The emergence of Silicon Valley and technological innovation vastly increased U.S. productivity. However, as the quote suggests, this period

---

[11] Taesuh Cha & Jungkun Seo, Trump by Nixon: Maverick Presidents in the Years of U.S. Relative Decline, 30 Korean J. Def. Analysis 79, 85, (2018).

[12] Quoted in Henry Kissinger, World Order (New York, NY: Penguin Press, 2015), 303.

[13] Mohammad Sarim, Realpolitik: The Pragmatic Approach to Politics, (June 4, 2025)

[14] The White House, A National Security Strategy of Engagement and Enlargement, 1, (1995).

brought not only promise but also uncertainty. First, although the collapse of the Soviet Union lead to a so-called "unipolar moment" for the United States, it also left the nation without a primary competitive target[15], forcing a shift in focus from a single rival to multiple issues. Second, while the Clinton administration's strong push for globalization boosted the economy, it simultaneously globalized risks. Finally, although the ensuing technological revolution massively boosted productivity, it also introduced a series of challenges, including cybersecurity threats, intellectual property protection, and high-tech export controls.

When summarizing the transformation of the national security concept, the most clear feature of this period was the formal inclusion of economic security as a component of national security. The existential threat driving this shift came from the globalization of the post-Cold War era—or, more accurately, the rapidly changing geopolitical landscape created by this globalization. With the collapse of the bipolar Cold War structure, focus shifted away from military preparation toward the globalized landscape and the fast-changing impact of economic changes on the global balance of power. Furthermore, under the dollar-based currency exchange mechanism of the Bretton Woods system, the United States had dominated the global economy for over two decades, during which economics never became a top-priority issue. However, with the breakdown of the Bretton Woods system, technological progress, and deepening global relations, economic security became a focal point of both foreign and domestic policy.[16] This shift toward a national security concept centered on economic security was reflected in the establishment of the National Economic Council[1718]. Its primary functions included coordinating the policy-making process for domestic and international economic issues, advising the President on economic policy, and monitoring the implementation of the President's econom-

---

[15] Alida Tomja, Internationalism of the Post-Cold War World: Clinton's Foreign Policy for the Global Age, 11 Interdisc. J. Rsch. & Dev. (2024).

[16] Chris J. Dolan & Jerel A. Rosati, U.S. Foreign Economic Policy and the Significance of the National Economic Council, 7 Int'l Stud. Persp. 102, (2006).

[17] Chris J. Dolan & Jerel A. Rosati, U.S. Foreign Economic Policy and the Significance of the National Economic Council, 7 Int'l Stud. Persp. 107, 103 (2006)

[18] Exec. Order No. 12,835, 58 Fed. Reg. 6,189 (Jan. 25, 1993) (establishing the National Economic Council).

ic agenda. In National Security For a New Strategy, national interests were divided into three categories.[19] The first category is core interests, which are broad and vital to the nation's survival, security, and vitality. These include territorial security, the safety of citizens, social and economic well-being, and the protection of critical infrastructure. The second category comprises important national interests. These are interests that, while not affecting the nation's survival, are crucial to its well-being and the structure of the world in which it exists. These include major economic interests, the protection of regions where the U.S. has obligations to allies, global environmental protection, and crises that could trigger significant impacts. The final category consists of humanitarian and other interests. Thus, it is clear that during the Clinton era, beyond establishing economic security as a core national security concern, there was an emerging trend toward broadening the concept of national security. The relevant issues expanded to include immigration, cybersecurity, and environmental protection.

1.4 The Fourth Transformation: 9/11 and the Shift to Preemption

On September 11, 2001, as the towering Twin Towers of the World Trade Center collapsed and planes struck the Pentagon—the symbol of America's highest military power—the fourth transformation of the U.S. national security concept officially took place. Without a doubt, the existential threat driving this shift was this unprecedented terrorist attack. The National Security Strategy of the United States of America stated: "Enemies in the past needed great armies and great industrial capabilities to endanger America. Now, shadowy networks of individuals can bring great chaos and suffering to our shores for less than it costs to purchase a single tank."[20] The securitized outcome of this transformation was a shift from "deterrence" to "preemption." The 9/11 attacks made the United States realize that terrorism was now its top national security priority. Because of the unique nature of this threat, its extreme destructiveness, disorder, and irrationality, the U.S. would be at a disadvantage in

[19] The White House, A National Security Strategy for a New Century, 1, (1999).

[20] The White House, The National Security Strategy of the United States of America, (2002).

this unpredictable war if it did not take the initiative. [21]"It has taken almost a decade for us to comprehend the true nature of this new threat. Given the goals of rogue states and terrorists, the United States can no longer solely rely on a reactive posture as we have in the past. The inability to deter a potential attacker, the immediacy of today's threats, and the magnitude of potential harm that could be caused by our adversaries' choice of weapons, do not permit that option. We cannot let our enemies strike first."[22] Regardless of the academic debates over the difference between "preventive war" and "preemption," and regardless of its legitimacy[23], this shift set the stage for the future development of national security. From this point forward, national security was no longer passive. It was no longer a shield waiting for a threat to strike before offering protection; instead, it became a sword, actively seeking out and eliminating potential threats.

Finally, during the Trump and Biden administrations, we can observe the fifth transformation of the national security concept. This era feels extraordinary—perhaps because we are living through its changes and developments firsthand, which highlights every detail. Following our previous framework for analyzing the national security concept, we will begin with the existential threat. This is an era of rapid change, where the world we face each day seems different from the last. I would summarize the driving force behind this as the development of dual-use technology. The most notable shift in this era is the disappearance of a clear boundary between military and civilian technologies. Data collected on platforms like short-video and dating apps can serve not only as the foundation for developing civilian industries, but can also—through the collection and analysis of big data—reflect the public opinion of a region or even an entire country. As a result, such data can be treated as state secrets, and its leakage could create a hidden risk to national security. The emergence of large AI language models has not only massively boosted the efficiency of private companies across various industries, but these models are also being ap-

---

[21] Shahid Ali Khattak, The Bush Doctrine of Pre emption and the US response after 9/11 attacks: Invasion of Afgha1nistan and Iraq, 18 J. Pol. Stud. 155, 157,(201).

[22] The White House, The National Security Strategy of the United States of America, 15, (2002).

[23] Colin S. Gray, The Implications of Preemptive and Preventive War Doctrines: A Reconsideration (Strategic Stud. Inst. 2007) (analyzing the legitimacy of the Bush Doctrine's preemptive strategy).

plied to military decision-making and operations. This has even created tension between the government and the private tech companies developing these advanced AI models. Growing geo-political is another key factor shaping the national security concept during this period. This escalation can largely be attributed to the technology born in this era; specifically, the borderless, intangible, and transferable nature of this technology has fueled the broader geopolitical competition. Furthermore, the pandemic played a critical role. During this unique period, the United States realized the vulnerability of its supply chains. In response to the backlash of globalization, it began to place a strong emphasis on domestic production.

1.5 The Fifth Transformation: Dual-Use Tech and the Weaponization of National Security

The securitized outcome of these existential threats, which is the transformation of the national security concept, can be seen in the weaponization of national security measures and the normalization of this weaponization. One manifestation of this shift is another focal point of this article: the tightening of investment regulations, which serves as a policy reflection of the broadening national security concept. As we will detail below, U.S. regulatory mechanisms for both inbound and outbound investment are no longer limited to mere oversight. Instead, they have taken on a highly targeted and offensive nature, turning into tools to win geopolitical competitions. Investment is just one aspect. The broadening of the national security concept in other areas also proves that using national security policy as an excuse to launch proactive competition and attacks has become an increasingly frequent and normal practice.

2. CURRENT LANDSCAPE OF U.S. INVESTMENT SCREENING MECHANISM

This chapter examines the U.S. investment screening mechanism from both inbound and outbound perspectives: the CFIUS framework and the OISP rules. The section on inbound screening traces the development, jurisdiction, and institutional structure of CFIUS. By analyzing Ralls Corp. v. CFIUS and the Nippon Steel case, it also highlights the practical challenges investors face when seeking legal remedies.

Finally, a comparison between the current U.S. system and the EU mechanism reveals the defining features of both approaches. The section on outbound screening first explains the necessity of regulating outbound investment and outlines its three primary review processes. Finally, a direct comparison of the inbound and outbound systems clarifies their unique characteristics, demonstrating how both mechanisms have increasingly become tools for geopolitical competition.

2.1 Inbound Investment Screening: The CFIUS Framework

CFIUS (Committee on Foreign Investment in the United States) was established by Executive Order 11858 signed by President Ford, with its legislative authority derived from the Defense Production Act[24]. It has three primary functions: first, analyzing investment trends in the U.S.; second, providing guidance for proposed major investment projects in consultation with foreign governments; and third, reviewing foreign investments that could significantly impact national interest and considering new legislative or regulatory proposals involving foreign investment. Thus, while the initial function of CFIUS was primarily to prevent major national security threats, its core objective remained the preservation of the principles of open investment and commercial freedom upheld by the United States. Entering the 1980s, the Fujitsu-Fairchild Semiconductor case prompted Congress to pass the Exon-Florio Amendment, granting the President the authority to block foreign transactions following a CFIUS review.[25] Next, the Foreign Investment and National Security Act of 2007 (FINSA)[26], driven by the "DP World" case, was enacted. This further expanded the definition of "national security," mandated review procedures, and increased Congressional oversight. FINSA established CFIUS's procedures for review and investigation, risk mitigation, tracking, and post-implementation monitoring.[27]

In 2018, President Trump signed the Foreign Investment Risk Review Moderni-

---

[24] Defense Production Act of 1950, 50 U.S.C. § 4565 (2024).

[25] Omnibus Trade and Competitiveness Act of 1988, Pub. L. No. 100-418, § 5021, 102 Stat. 1107 (1988)

[26] Foreign Investment and National Security Act of 2007, Pub. L. No. 110-49, 121 Stat. 246 (2007).

[27] Ibid., § 2 (codified at 50 U.S.C. § 4565.

zation Act (FIRRMA)[28], introducing a mandatory declaration system for investments involving foreign government interests, shifting the entire mechanism toward a more targeted and offensive posture. In September 2022, President Biden signed an Executive Order[29] requiring CFIUS to enhance scrutiny of foreign investments in sectors such as cybersecurity, quantum computing, biotechnology, and sensitive data. Meanwhile, the "America First" policies of the Trump administration specifically listed China, Russia, Iran, North Korea, Cuba, and Venezuela as "foreign adversaries" , transitioning the review mechanism from a defensive to an offensive mode.

In examining the current mechanism, the core responsibility of CFIUS—as stated in the CFIUS 2024 Annual Report to Congress—is "to address any national security risk that arises as a result of such transactions."[30] First, the scope of "transactions involving national security" is extremely broad. This is due to the overly broad current U.S. definition of national security, where technological sensitivity, cross-border capital flows, and geopolitical factors have expanded the review scope. In modern society, characterized by rapid technological innovation, CFIUS's reach into all areas of the digital economy—coupled with a lack of clear definitions—risks obstructing investment and creating significant uncertainty for investors. Furthermore, the decision-making process involves multiple agencies, including the Treasury, State, Defense, Justice, Commerce, Energy, and Homeland Security; as well as the USTR and the OSTP within the Executive Office of the President.[31] The involvement of numerous agencies increases the number of variables considered, further complicating the review process.[32]

Regarding the review mechanism, the U.S. employs a "dual-track" system consist-

---

[28] Foreign Investment Risk Review Modernization Act of 2018, Pub. L. No. 115-232, tit. XVII, 132 Stat. 2173 (2018).

[29] Exec. Order No. 14083, 87 Fed. Reg. 57369 (Sept. 15, 2022).

[30] U.S. Department of the Treasury, Committee on Foreign Investment in the United States: Annual Report to Congress for Calendar Year 2024 (Washington, DC: 2025), 5.

[31] See Exec. Order No. 11,858, as amended.

[32] Floor Doppen, Antonio Calcara, and Dirk De Bièvre, "Balancing Security and Economics: Domestic State-firm Relations and Investment Screening Mechanisms in Europe," Review of International Political Economy 31, no. 6 (2024): 1641-42.

ing of voluntary and mandatory declarations.[33] On the surface, this model appears selective outside of mandatory triggers; however, in practice, this is not the case. If the Committee has reason to believe a transaction may fall under CFIUS jurisdiction and raise national security concerns, it possesses the authority to review pending or completed transactions even if the parties did not file voluntarily. The CFIUS final regulations effective December 2024 significantly increased the maximum penalty for a single violation from $250,000 to $5 million.[34] Beyond heavy fines, investors face the risk of forced divestment and transaction rescission even after a deal has closed.

Remedies available to investors are highly limited. First, investors may enter into Mitigation Agreements with the government, designed to reduce national security risks through measures such as divesting sensitive technology assets or restricting access rights.[35] Second, some investors have sued the government and the President based on due process reviews. Two notable cases in this regard are Ralls Corp. v. CFIUS[36] and the Nippon Steel case.

The Ralls case involved a Chinese-owned enterprise acquiring four small companies to develop wind farms near a military base in Oregon. CFIUS ordered Ralls to sell the companies, remove all construction, and vacate the site. President Obama personally issued a similar order requiring divestment, claiming credible evidence that the action might threaten to impair the national security of the United States. Ralls sued CFIUS and the President in the U.S. District Court for the District of Columbia (D.D.C.). Despite government opposition, Judge Amy Berman Jackson ruled the case justiciable but dismissed it on two grounds: the claim against CFIUS was mooted by the Presidential action, and Ralls was deemed to have waived its due process rights by failing to notify CFIUS prior to the acquisition. Ralls appealed.[37] The key issue in the second instance was whether due process could be reviewed,

---

33 31 C.F.R. §§ 800.401, 800.501 (2024).

34 Penalty Provisions, Provision of Information, Negotiation of Mitigation Agreements, and Other Procedures Pertaining to Certain Investments in the United States by Foreign Persons, 89 Fed. Reg. 93179 (Nov. 26, 2024) (codified at 31 C.F.R. pts. 800, 802).

35 50 U.S.C. § 4565(l).

36 Ralls Corp. v. Comm. on Foreign Inv. in the U.S., 758 F.3d 296 (D.C. Cir. 2014).

37 Ralls Corp. v. Comm. on Foreign Inv. in U.S., 758 F.3d 296 (D.C. Cir. 2014).

even if the substantive decision itself was unreviewable under Presidential authority. The court ultimately held that the "political question doctrine did not bar court from deciding whether Due Process Clause entitled corporation to have notice of, and access to, evidence on which President relied and opportunity to rebut that evidence before he reached his non-justiciable and statutorily unreviewable divestiture determination". [38]On one hand, the Ralls appeal significantly increased the transparency of CFIUS reviews. In 2018, FIRRMA substantially amended Section 721, adding provisions allowing for judicial review. FIRRMA granted the D.C. Circuit original and exclusive jurisdiction over civil actions challenging Presidential actions or findings.[39] As a result, investors can learn why an investment was terminated, gain a concept of CFIUS standards, and adjust future investment directions. On the other hand, the final judgment did not provide Ralls with a substantive remedy. While it allowed for the review of due process, it only ensured investors know why a decision was made, without any right to challenge the decision itself. This obstacle arises from Section 721 of the Defense Production Act of 1950, which states that "The actions of the President under paragraph (1) of subsection (d) and the findings of the President under paragraph (4) of subsection (d) shall not be subject to judicial review." [40]Thus, investors have no right to overturn decisions, rendering the remedies insubstantial.

In the Nippon Steel case, Japan's Nippon Steel Corp proposed to acquire U.S. Steel for approximately $14,900 million in cash. Despite Japan being a close ally, the deal faced strong opposition from the United Steelworkers (USW). Thus, Nippon Steel pledged to protect union benefits and increase investment, leading to a merger agreement. Subsequently, Biden publicly opposed the deal, and CFIUS initiated a review. During this period, Nippon Steel repeatedly attempted to submit "mitigation agreements" but received no substantive concessions. In 2025, President Biden formally issued an Executive Order blocking the acquisition on national security grounds. Nippon Steel then filed suit in the D.D.C. challenging the order. The Executive Order claimed Japan might take action that threatens to impair the national security of the

---

[38] Ralls Corp. v. Comm. on Foreign Inv. in U.S., 758 F.3d 296 (D.C. Cir. 2014).

[39] Hdeel Abdelhady, "The Nippon-U.S. Steel Case Tests Presidential Power to Block Foreign Investment, Judicial Review," MassPoint Legal and Strategy Advisory, January 2, 2025. For the statutory basis, see 50 U.S.C. § 4565(p), as added by Section 1715 of FIRRMA (2018).

[40] 50 U.S.C. § 4565.

United States, without specifying the threat. While withholding classified details is common, the order also mentioned "to ensure that, now and long into the future, America has a strong domestically owned and operated steel industry" and emphasized that "U.S. Steel will remain a proud American company... American-owned, American-operated, by American union steelworkers - the best in the world."[41] This phrasing is the main dispute. The questions to analyze are: first, while the Act grants the President discretion, the statute conferring power to CFIUS and the President enumerates specific factors for national security assessment—is this list merely illustrative or substantively restrictive?[42] In other words, where is the line between the President's subjective judgment and an objective determination of national security? Second, the President cited unfair trade practices, domestic steel demand, and job losses as grounds for the order, thereby incorporating unions and employment into the scope of "national security," which clearly exceeds the boundaries set by Congress.[43] FINSA (2007) specifies that "national security" encompasses issues relating to homeland security, including its application to critical infrastructure, and Congress has repeatedly refused to expand the scope of Section 721.[44] On the day the order was signed, the Treasury Secretary and CFIUS Chair issued a statement claiming the order merely provided "context" and "did not change the CFIUS process."[45] Therefore, is President Biden's decision in jeopardy for involving unauthorized acts beyond Congressional mandate?

Comparing the current U.S. investment review status with the EU mechanism reveals several characteristics. First, the U.S. mechanism has mandatory enforcement power. The "Final Rule" issued by the Treasury explicitly outlining penalties for

---

41 Hdeel Abdelhady, The Nippon-U.S. Steel Case Tests Presidential Power to Block Foreign Investment, Judicial Review, MassPoint PLLC (2025), https://masspointpllc.com/.

42 50 U.S.C. § 4565(f).

43 Hdeel Abdelhady, "The Nippon-U.S. Steel Case Tests Presidential Power to Block Foreign Investment, Judicial Review," MassPoint Legal and Strategy Advisory, January 2, 2025. For the statutory basis, see 50 U.S.C. § 4565(p), as added by Section 1715 of FIRRMA (2018).

44 Foreign Investment and National Security Act of 2007, Pub. L. No. 110-49, 121 Stat. 246 (2007).

45 Press Release, U.S. Dep't of the Treasury, Statement by Secretary of the Treasury Janet L. Yellen on President Biden's Executive Order on the Committee on Foreign Investment in the United States (Sept. 15, 2022), https://home.treasury.gov/news/press-releases/jy0951.

violations. In contrast, the EU mechanism is largely at a recommendation stage; the EU does not mandate that every member state establish its own review mechanism, and the final decision on whether a specific investment should be blocked remains with individual states. EU Commission recommendations are not legally binding but serve as tools for guidance and policy coordination. Second, U.S. reviews are clearly targeted. The U.S. mechanism lists China, Russia, Iran, North Korea, Cuba, and Venezuela as foreign adversaries, providing a clear focus. The EU maintains national neutrality, and its recommendations do not name specific countries, urging member states to assess based on their own risk profiles. Finally, the U.S. logic emphasizes active intervention and resolution, directly prohibiting transactions deemed significant threats, while requiring notification for others. The EU focus remains on information gathering and risk assessment to guide subsequent decision-making.

It is worth noting that this article does not advocate for a completely open or unrestricted policy toward Foreign Direct Investment (FDI). In the current era, FDI does carry risks that need prevention and reflection, such as supply chain dominance and technology transfer that may harm host country interests. Investments in Artificial Intelligence are particularly prone to risks, including unauthorized data leaks, foreign adversaries gaining access to sensitive tech or facilities, the introduction of supply chain vulnerabilities, and intellectual property theft. Therefore, the view here is not that investment review mechanisms are redundant or lack legitimacy, but rather to point out that current reviews have, to an extent, exceeded their original mandate, transforming from risk prevention tools into instruments of geopolitical competition.

2.2 Outbound Investment Screening: The OISP Mechanism

In May 2021, Senators John Cornyn and Bob Casey first introduced the National Critical Capabilities Defense Act (NCCDA), a rare bill supported by both parties.[46] This proposal aimed to establish an outbound investment screening mechanism,

[46] N. Zamani, A Legal Comparative Approach towards the Screening of Outbound FDI. What Can the EU and Its Member States Learn from the US National Critical Capabilities and Defense Act Proposal?, 2022 Erasmus L. Rev. 1, 1, (2022).

requiring relevant agencies to monitor critical supply chains and submit annual reports to Congress. The House of Representatives incorporated the NCCDA provisions into the America COMPETES Act of 2022. Its screening scope was exceedingly broad, even encompassing regulations on general industry offshoring. In the summer of 2022, the U.S. Chamber of Commerce, Wall Street, and multinational tech giants joined forces to lobby to oppose the bill. Ultimately, Congress completely removed the NCCDA, and the legislation failed. However, on August 9, 2023, President Biden signed Executive Order 14105, which narrowed the scope of the NCCDA to focus on three core technologies: artificial intelligence, semiconductors, and quantum technology. On October 28, 2024, the U.S. Department of the Treasury officially released the final implementing regulations, known as 31 CFR Part 850.[47] Notably, both Executive Order 14105 and 31 CFR Part 850 establish an explicitly discriminatory stance against China, using geographic and identity discrimination as geopolitical weapons.

The tightening of investment regulations is a well-established trend. Previous sections of this paper have extensively examined this phenomenon, detailing the scope and impact of CFIUS's oversight on inbound investments. However, the regulation of outbound direct investments seems to have received less public attention.[48] Why did the U.S. persist in using a Presidential Executive Order to regulate outbound investment after the bill failed? Why is outbound investment just as important as inbound investment? The context behind this phenomenon is, in fact, similar to that of inbound investment regulation: driven by geopolitics and the pandemic, the United States realized the vulnerability of its supply chains in a post-globalization world.[49] If U.S. major companies, seeking commercial profits, constantly pour capital overseas to build large-scale production lines in other countries, it will lead to a loss of related manufacturing capabilities back in United States. This may not seem like a

---

47 Brooks E. Allen, Katie Clarke, Brian J. Egan & Michael E. Leiter, Congress Reportedly Advances Broad Proposal for Outbound Screening of US Investments in Identified Countries of Concern, Including China, Skadden (June 21, 2022)

48 N. Zamani, A Legal Comparative Approach towards the Screening of Outbound FDI. What Can the EU and Its Member States Learn from the US National Critical Capabilities and Defense Act Proposal?, 2022 Erasmus L. Rev. 1, 1, (2022).

49 Jonas Fechter, Next-level Screening? The Case of Outbound Investment Screening

problem in the short term. However, if a severe geopolitical conflict occurs, an opposing nation could ban exports and completely cut off U.S. supply chains. Furthermore, outbound investment involves not just money, but also the technology, management, and experience that comes with it. Investors help these foreign enterprises establish governance structures matching their own and integrate them into their own supply chains. Therefore, the U.S. worries that outbound investment in specific sectors will instead leak the nation's critical technologies and industrial expertise, unintentionally helping the technological development of other nations in a highly competitive geopolitical arena where the gap between countries is not significantly wide. Naturally, another reason for regulating outbound direct investment is to prevent the evasion of export controls,[50] stopping investors from exporting advanced technologies and products, which are prohibited from being sold to other nations under the Export Control Reform Act, through the channel of outbound investment. From the analysis above, it is evident that the regulation of outbound investment has a reasonable necessity. However, the tension between its practical implementation and market freedom is another issue we must analyze.

Because the National Critical Capabilities Defense Act (NCCDA) failed to pass, National Critical Capabilities Committee was never established. In the final regulations officially released by the U.S. Department of the Treasury, 31 CFR Part 850 (the OISP rules), the operational responsibility for this "reverse CFIUS"[51] mechanism did not fall to any newly created committee. Instead, the Treasury Department serves as the core leading agency, consulting with the Departments of Commerce, State, Defense, Energy, and the Office of the Director of National Intelligence when drafting rules, assessing national security threats, and processing exemptions[52]. The main process of this reverse CFIUS mechanism relies primarily on compliance self-assessments, online notifications, and official reviews. First, the Treasury Department

---

[50] N. Zamani, A Legal Comparative Approach towards the Screening of Outbound FDI. What Can the EU and Its Member States Learn from the US National Critical Capabilities and Defense Act Proposal?, 2022 Erasmus L. Rev. 1, 1, (2022).

[51] See, e.g., President Biden Issues Executive Order Restricting Outbound Investment in National Security Technologies and Products, 117 AM. J. INT'L L. 713, 715 (2023).

[52] Provisions Pertaining to U.S. Investments in Certain National Security Technologies and Products in Countries of Concern, 89 Fed. Reg. 90,398 (Nov. 15, 2024)

defines a U.S. person's "knowledge" regarding a transaction as "actual knowledge that a fact or circumstance existed or was substantially certain to occur, an awareness of a high probability of a fact or circumstance's existence or future occurrence, or reason to know of a fact or circumstance's existence." Furthermore, investors are obligated to conduct a "reasonable and diligent inquiry." Notably, the Treasury declined to include a safe harbor provision in the final rules.[53] A safe harbor provision would mean that if investors could prove their actions fully met the pre-set standards or compliance processes established by regulators, they would be legally exempt from specific liabilities, administrative penalties, or criminal charges. Instead, the OISP rules aim to provide a non-exhaustive list of considerations to address the highly fact-specific and individualized nature of each transaction. Second, regarding the online notification process, U.S. investors must submit their notifications to the Treasury no later than 30 calendar days after the completion date of a transaction. Finally, individuals who violate, attempt to violate, or conspire to violate these rules may face maximum civil penalties. The penalty amount will be the greater of twice the value of the transaction that constitutes the violation or $250,000. Cases involving individuals whom the Treasury believes have committed potential criminal violations may be referred to the Attorney General.[54]

Having outlined the basic mechanisms of CFIUS and the reverse CFIUS (OISP), let us compare the two.[55] First, CFIUS and OISP serve different regulatory purposes. When investors file with CFIUS, their primary goal is to secure a "safe harbor." They do this to prevent the government from later imposing penalties or forcing a divestment based on national security concerns. Under the OISP rules, however, investors must file simply because their transaction involves covered activities; failing to notify the government, or investing in explicitly prohibited sectors, triggers strict penalties. Unlike CFIUS, an OISP filing does not grant investors a safe harbor. Even after submitting a notification, the Treasury can still penalize investors if it finds any procedural violations. Furthermore, the Treasury requires U.S. persons to update

53 Id.

54 Id.

55 Harlan Grant Cohen, Outbound Investment Restrictions and International Law's Challenge, 49 Seattle U. L. Rev.

their notifications on an ongoing basis. This rule has sparked widespread debate in the legal community. Critics argue that this updating obligation should only apply when an investor is considering a follow-on investment in the target company, asserting that investment regulations should not become a tool for the U.S. government to monitor specific targets.[56] Second, investors bear different responsibilities under the CFIUS and OISP mechanisms.In CFIUS, while investors should review their own investment before applying, once the case is submitted, the Committee bears the responsibility of gathering information and conducting the investigation. Under the OISP, investors bear a heavier screening burden; they must fulfill the obligation of a "reasonable and diligent inquiry" and file a notification only after confirming that a covered issue exists. Third, a major characteristic of the CFIUS process is its ability to negotiate mitigation agreements with investors on a case-by-case basis. While the OISP clearly states that U.S. persons may apply for an exemption from the disclosure obligations or relevant prohibitions outlined in the Final Rule, the Treasury declined to specify the exact conditions that transactions must meet to qualify for a national interest exemption.[57] Furthermore, if an exemption request is denied, the Treasury currently has no established appeal process or time frame. Thus, the logic of CFIUS is that if a transaction poses a national security risk, the Committee can negotiate with the company to mitigate that risk and formulate a plan. Under the OISP rules, however, the outcome is binary: a transaction is either exempted or prohibited, leaving no room for negotiation or modification. As a result, some scholars argue that CFIUS[58], because of its efforts to allow investments to succeed, is viewed as a mechanism designed to promote economic development while safeguarding national security. In contrast, the OISP, lacking flexibility and making no effort to facilitate investments, is viewed as a "sanctions regime, prohibiting and criminalizing certain transactions."[59] Rather than merely safeguarding na-

---

[56] U.S.-China Bus. Council, USCBC Comments Regarding the Proposed Rule Pertaining to US Investments in Certain National Security Technologies and Products in Countries of Concern

[57] Provisions Pertaining to U.S. Investments in Certain National Security Technologies and Products in Countries of Concern, 89 Fed. Reg. 90,398 (Nov. 15, 2024)

[58] Harlan Grant Cohen, Outbound Investment Restrictions and International Law's Challenge, 49 Seattle U. L. Rev.

[59] Id.

tional security, its objective seems more focused on containing geopolitical and strategic rivals, further demonstrating that U.S. regulatory policy is becoming less defensive and more offensive.

3. ANALYSIS OF THE DRIVERS BEHIND THE TIGHTENING OF AI INVESTMENT SCRUTINY

The drivers behind the tightening of investment scrutiny regarding Artificial Intelligence (AI) are analyzed from two perspectives: first, how the characteristics of AI trigger the tightening of review mechanisms; and second, how the external environment further accelerates this trend.

From the perspective of AI's nature, the characteristics of "dual-use" and its dominance by the private sector have largely driven regulatory intensification. First, the blurring of the "dual-use" boundary is a critical factor. Traditionally, "dual-use" refers to technology serving both civilian and military purposes. For instance, while Large Language Models (LLMs) significantly enhance productivity in the civilian sector, their underlying algorithms can be "weaponized" in a military context for autonomous combat systems or cyber warfare.[60] This dual capability makes AI a focal point of national security risk assessments. In early 2026, OpenAI signed an agreement with the U.S. Department of Defense, permitting its AI technology for military and security operations.[61] This decision promptly drew heavy criticism on social media and sparked a "#CancelChatGPT" movement. Notably, many ChatGPT users switched to Claude after it publicly rejected a similar Pentagon request. Both ChatGPT and Claude are large-scale language models and the most widely used public AI tools. Daily experience suggests these models are harmless; however, this "harmlessness" lies in how it is used, not what it is. LLMs, machine learning, and other AI technologies provide strategic, operational, and tactical planning support to

[60] Kelley M. Sayler, Artificial Intelligence and National Security, Cong. Rsch. Serv., R45178, at 14–16 (2023)

[61] Theo Farrant, 'Cancel ChatGPT': AI Boycott Surges After OpenAI-Pentagon Military Deal, Euronews (Mar. 2, 2026), https://www.euronews.com/next/2026/03/02/cancel-chatgpt-ai-boycott-surges-after-openai-pentagon-military-deal.

military decision-makers by rapidly integrating fragmented data.[62] Given that AI still possesses various flaws, its application at the military level risks causing casualties among innocent civilians. Second, innovation dominated by the private sector has created national security anxieties, leading to institutional compensation and regulatory challenges.[63] In the current technological landscape, AI progress is primarily driven by private giants such as OpenAI, Microsoft, and Google. This shift shows a relative weakening of direct government control over strategic technologies. Feeling this loss of control, governments often react by enforcing stricter investment reviews.

From the perspective of the external environment, multiple factors have combined to speed up this tightening. First, with the rapid evolution of massive datasets and AI, the economic center of gravity is shifting from tangible assets to intangible intellectual property, such as algorithms and patents. Intangible assets are highly elusive and mobile; in the AI field, their core is often embodied in code and top talent. Once such assets are transferred via cross-border M&A, they are difficult to identify or recover quickly. This characteristic forces regulators to rely more heavily on upfront reviews to prevent irreversible risks. Second, there is a trend toward the "weaponization" of global supply chains. AI development depends heavily on computing power and data resources; if foreign capital controls key nodes in the supply chain, it may trigger risks such as "supply blackmail."[64] Surprisingly, the tightening of review mechanisms globally in recent years coincided with the COVID-19 pandemic. Theoretically, during a global economic contraction, investment reviews should be open. [65] Instead, despite the slow post-pandemic recovery, jurisdictions have tightened scrutiny, with supply chain panic being a core factor.[66] The pandemic made nations

---

[62] Kelley M. Sayler, Artificial Intelligence and National Security, Cong. Rsch. Serv., R45178, at 14–16 (2023)

[63] Floor Doppen, Antonio Calcara & Dirk De Bièvre, Balancing Security and Economics: Domestic State-Firm Relations and Investment Screening Mechanisms in Europe, 31 Rev. Int'l Pol. Econ. 1637, (2024).

[64] Mike Driscoll, Joe Knight & Eric Rudolph, Navigating National Security Risks in AI-Related Investments, FTI Consulting (Oct. 8, 2025)

[65] Mennill, S. (2014). Competing for capital: The political economy of foreign direct investment inflows and domestic employment. International Organization, 68(2), 295–321.

[66] Davis, C. L., & Pelc, K. (2017). Cooperation in hard times: Self-restraint of trade protection. Journal of Conflict Resolution, 61(2), 398–429.

realize that the long-pursued "globalization of supply chains" is extremely fragile. In the event of natural disasters or war, supply chain disruptions can deal a devastating blow to a nation. Furthermore, intensified geoeconomic competition and the rise of emerging economies, such as China, have transformed investment review mechanisms into tools for strategic competition. As de-globalization deepens, countries are using stricter review systems to selectively detach from global networks, thereby securing their own security. Specifically, driven by the weaponization of national security concepts, the international investment regulatory framework has increasingly become an instrument of geopolitical warfare. Both Executive Order 14105 and 31 CFR Part 850 contain explicitly discriminatory, country-specific provisions. As a prime example within the OISP framework, the Treasury Department mandates that investments in entities from a "country of concern" are prohibited or subject to notification even if their involvement in regulated activities is minimal. This shows a clear offensive strategy in the tech war, prioritized over U.S. economic interests.

4. ASSESSMENT OF THE TIGHTENED U.S. INVESTMENT REVIEW SYSTEM

For inbound investors, first, agencies can still review and reverse transactions that have already been completed. While filing is technically voluntary, failing to do so means harsh penalties. The threat of substantial fines, along with mandated divestment and transaction rescission, transforms what was once a choice for investors into an unavoidable necessity. Second, the U.S. review mechanism has been criticized as a "black box" operation, the lack of transparency in the review process undermines the reasonable expectations of investors.[67] Furthermore, the ambiguity of review standards makes it difficult for investors to determine how to comply, creating a "chilling effect" on investment.[68] While the United States has historically prided itself on being the most liberal economy for investment, attracting hundreds of billions of

[67] Jonathan Wakely & Andrew Indorf, Managing National Security Risk in an Open Economy: Reforming the Committee on Foreign Investment in the United States, 9 Harv. Nat'l Sec. J. 1 (2018)

[68] Edmond, C. "Strategic Knowledge: Innovation, Technological Sovereignty, and Geopolitical Competition." Knowledge Exposition, 2026.

dollars in foreign capital annually as a premier investor and recipient, the recent tightening of review mechanisms has caused investors to hesitate due to high risks of massive compensation and investment suspension. States like Louisiana, whose economies are heavily dependent on energy production, maritime trade, petrochemicals, and advanced manufacturing, are more heavily affected. As investors shift to regions with lower regulatory barriers, U.S. competitiveness in the international market is compromised. Third, investors lack effective judicial remedies. Current options are limited to either entering into a risk mitigation agreement with the government, which may result in investment losses, or pursuing litigation regarding due process, a path that remains highly controversial and fails to provide substantive relief. For outbound investors, First, investors face a heavy due diligence burden. They must conduct a "reasonable and diligent inquiry" under an exceptionally broad standard of "knowledge"—deemed to exist both where actual knowledge exists and where there is a “high probability” that the target company is engaged in covered activities. Such overwhelming obligations often force investors to abandon transactions entirely. Second, investors remain legally vulnerable. Even if a notification is completed within the required 30 days post-transaction, the Treasury retains the authority to penalize any procedural violations. More severely, the obligation to "update notifications on an ongoing basis" effectively transforms investors into government monitoring tools, trapping them in a perpetual compliance loop. Third, the OISP's exemption rules are highly unclear. The national interest exemption lacks critical operational details, such as specific application requirements and explicit approval timelines, leaving investors with no clear expectations. Finally, beyond strict compliance demands, investors face severe legal consequences. Any violation, attempted violation, or conspiracy to violate the rules triggers heavy civil fines, and cases deemed potential criminal violations by the Treasury are referred to the Attorney General for prosecution.

On the international stage, criticism of the U.S. mechanism is growing. First, CFIUS is increasingly becoming a geopolitical tool, gradually restricting investment freedom. In the tension between national security and investment freedom, the "America First" policy[69] has signaled a clear preference. The mandate of CFIUS is no

[69] The White House, National Security Strategy of the United States of America (Dec. 2017)

longer just to balance interests, but to prioritize a vaguely defined and broad "national security" over investment freedom. Official "America First" documents repeatedly cite threats from hostile nations, specifically targeting China. In other words, "identity" has become a threshold factor; if the "identity" is that of a hostile nation, transaction face much tighter security. This specific targeting shows that the U.S. system is no longer just defensive; it has become actively interventionist and even punitive. If identity becomes the deciding factor rather than a secondary consideration, affected parties could argue that CFIUS, as an executive body, has exceeded the authority granted by FIRRMA. Second, as a nation of significant influence, the U.S. policy of tightening reviews accelerates de-globalization and surges neo-protectionism.[70] This disrupts global AI cooperation, fragmenting global governance and making it difficult to reach a regulatory consensus based on the common interests of humanity. Regarding innovation, technology is no longer borderless, global innovation chains are severed by geopolitical boundaries, and ultimately slows down innovation.

CONCLUSION

The evolution of the U.S. national security concept reveals a clear trend of expansion and weaponization. Historically, this concept has undergone five distinct transformations: shifting to long-term peacetime preparation during the Cold War; prioritizing a balance of power over ideology in the Nixon era; including economic security during the Clinton administration; adopting a preemptive, offensive posture after 9/11; and finally, weaponizing security measures in today's era of big data and artificial intelligence (AI). Through these shifts, the definition of national security has become increasingly broad and blurred, allowing it to deeply influence and shape concrete policy making.

This expansion directly drives the current tightening of U.S. investment regulations. On the inbound side, the CFIUS framework exercises broad authority through a dual-track system. As the *Ralls* and *Nippon Steel* cases illustrate, investors lack effective legal remedies, highlighting the highly controversial line between a Pres-

[70] Julian Hinz et al., Mild Deglobalization: Foreign Investment Screening and Cross-Border Investment, Kiel Inst. for the World Econ., Working Paper No. 2265, at 5-6 (2024)

ident's subjective political judgment and objective security risks. On the outbound side, the OISP rules (31 CFR Part 850) enforce strict self-assessment, notification, and Treasury review requirements. Comparing these two mechanisms shows that U.S. investment screening has departed from its original purpose of risk prevention, becoming instead a strategic weapon for geopolitical competition.

The tightened scrutiny of AI investments clearly illustrates this trend. While the dual-use nature of AI makes some level of oversight necessary, other factors drive the current regulatory extreme. Because the private sector dominates advanced AI research, the government feels a loss of control and uses regulation to reclaim it. Furthermore, the global transition to intangible intellectual property, caused by pandemic supply chain panics, and increasing geopolitical tensions have all accelerated the push for stricter AI oversight.

Ultimately, this weaponization of investment policy carries severe consequences. For investors, unclear review processes and the lack of judicial relief create a serious chilling effect, blocking capital from critical technological sectors. On the international stage, this regulatory tightening acts as an institutional tool for geopolitical dominance. It triggers a global chain reaction, fragmenting the international community and severely undermining efforts to establish cooperative global AI governance.

REFERENCES

31 C.F.R. §§ 800.401, 800.501 (2024).

Alida Tomja, Internationalism of the Post-Cold War World: Clinton's Foreign Policy for the Global Age, 11 Interdisc. J. Rsch. & Dev. (2024).

Bear F. Braumoeller, The Myth of American Isolationism, 6 Foreign Pol'y Analysis 349 (2010).

Brooks E. Allen, Katie Clarke, Brian J. Egan & Michael E. Leiter, Congress Reportedly Advances Broad Proposal for Outbound Screening of US Investments in Identified Countries of Concern, Including China, Skadden (June 21, 2022).

Chris J. Dolan & Jerel A. Rosati, U.S. Foreign Economic Policy and the Significance of the National Economic Council, 7 Int'l Stud. Persp. 102 (2006).

Colin S. Gray, The Implications of Preemptive and Preventive War Doctrines: A Reconsideration (Strategic Stud. Inst. 2007).

David A. Baldwin, quoted in Barry Buzan & Lene Hansen, The Evolution of International Security

Studies 67 (2009).

Davis, C. L., & Pelc, K. (2017). Cooperation in hard times: Self-restraint of trade protection. Journal of Conflict Resolution, 61(2), 398–429.

Defense Production Act of 1950, 50 U.S.C. § 4565 (2024).

Edmond, C. "Strategic Knowledge: Innovation, Technological Sovereignty, and Geopolitical Competition." Knowledge Exposition, 2026.

Exec. Order No. 11,858, as amended.

Exec. Order No. 12,835, 58 Fed. Reg. 6,189 (Jan. 25, 1993).

Exec. Order No. 14,083, 87 Fed. Reg. 57,369 (Sept. 15, 2022).

Floor Doppen, Antonio Calcara & Dirk De Bièvre, Balancing Security and Economics: Domestic State-Firm Relations and Investment Screening Mechanisms in Europe, 31 Rev. Int'l Pol. Econ. 1637 (2024).

Foreign Investment and National Security Act of 2007, Pub. L. No. 110-49, 121 Stat. 246 (2007).

Foreign Investment Risk Review Modernization Act of 2018, Pub. L. No. 115-232, tit. XVII, 132 Stat. 2173 (2018).

Harlan Grant Cohen, Outbound Investment Restrictions and International Law's Challenge, 49 Seattle U. L. Rev.

Hdeel Abdelhady, The Nippon-U.S. Steel Case Tests Presidential Power to Block Foreign Investment, Judicial Review, MassPoint Legal and Strategy Advisory (Jan. 2, 2025), https://masspoint-pllc.com/.

Jonas Fechter, Next-level Screening? The Case of Outbound Investment Screening.

Jonathan Wakely & Andrew Indorf, Managing National Security Risk in an Open Economy: Reforming the Committee on Foreign Investment in the United States, 9 Harv. Nat'l Sec. J. 1 (2018).

Julian Hinz et al., Mild Deglobalization: Foreign Investment Screening and Cross-Border Investment, Kiel Inst. for the World Econ., Working Paper No. 2265 (2024).

Kelley M. Sayler, Artificial Intelligence and National Security, Cong. Rsch. Serv., R45178 (2023).

Mennill, S. (2014). Competing for capital: The political economy of foreign direct investment inflows and domestic employment. International Organization, 68(2), 295–321.

Mike Driscoll, Joe Knight & Eric Rudolph, Navigating National Security Risks in AI-Related Investments, FTI Consulting (Oct. 8, 2025).

Mohammad Sarim, Realpolitik: The Pragmatic Approach to Politics (June 4, 2025).

Nat'l Sec. Council, NSC 68: United States Objectives and Programs for National Security (Apr. 14, 1950).

National Security Act of 1947, 50 U.S.C. § 3021(b).

National Security Act of 1947, Pub. L. No. 80-253, 61 Stat. 495.

Nicholas Steneck, National Security Act of 1947, in Routledge Resources Online – The Cold War (Ruud van Dijk & Liliane Stadler eds., 2025).

N. Zamani, A Legal Comparative Approach towards the Screening of Outbound FDI: What Can the EU and Its Member States Learn from the US National Critical Capabilities and Defense Act Proposal?, 2022 Erasmus L. Rev. 1 (2022).

Omnibus Trade and Competitiveness Act of 1988, Pub. L. No. 100-418, § 5021, 102 Stat. 1107 (1988).

Penalty Provisions, Provision of Information, Negotiation of Mitigation Agreements, and Other Procedures Pertaining to Certain Investments in the United States by Foreign Persons, 89 Fed. Reg. 93,179 (Nov. 26, 2024).

President Biden Issues Executive Order Restricting Outbound Investment in National Security Technologies and Products, 117 Am. J. Int'l L. 713 (2023).

Press Release, U.S. Dep't of the Treasury, Statement by Secretary of the Treasury Janet L. Yellen on President Biden's Executive Order on the Committee on Foreign Investment in the United States (Sept. 15, 2022), https://home.treasury.gov/news/press-releases/jy0951.

Provisions Pertaining to U.S. Investments in Certain National Security Technologies and Products in Countries of Concern, 89 Fed. Reg. 90,398 (Nov. 15, 2024).

Quoted in Henry Kissinger, World Order (New York, NY: Penguin Press, 2015), 303.

Ralls Corp. v. Comm. on Foreign Inv. in the U.S., 758 F.3d 296 (D.C. Cir. 2014).

Richard M. Jennings, The Thrust of the Nixon Doctrine (1972).

Robert D. Russell II, The National Security Debate and the Truman Administration's Policy Toward China, 1947-1950 (2000).

Shahid Ali Khattak, The Bush Doctrine of Preemption and the US Response after 9/11 Attacks: Invasion of Afghanistan and Iraq, 18 J. Pol. Stud. 155, 157 (201).

Taesuh Cha & Jungkun Seo, Trump by Nixon: Maverick Presidents in the Years of U.S. Relative Decline, 30 Korean J. Def. Analysis 79 (2018).

The White House, A National Security Strategy for a New Century (1999).

The White House, A National Security Strategy of Engagement and Enlargement (1995).

The White House, National Security Strategy of the United States of America (Dec. 2017).

The White House, The National Security Strategy of the United States of America (2002).

Theo Farrant, 'Cancel ChatGPT': AI Boycott Surges After OpenAI-Pentagon Military Deal, Euronews (Mar. 2, 2026), https://www.euronews.com/next/2026/03/02/cancel-chatgpt-ai-boycott-surges-after-openai-pentagon-military-deal.

U.S. Department of the Treasury, Committee on Foreign Investment in the United States: Annual

Report to Congress for Calendar Year 2024 (Washington, DC: 2025).

U.S.-China Bus. Council, USCBC Comments Regarding the Proposed Rule Pertaining to US Investments in Certain National Security Technologies and Products in Countries of Concern.

Walter Lippmann, U.S. Foreign Policy: Shield of the Republic 9–10 (1943).

ALGORITHMIC CONSTITUTIONALISM: FROM THE GUARANTEE TO THE RECONFIGURATION OF FUNDAMENTAL RIGHTS

Dra. Laura Nahabetián Brunet, PhD
Tenured Associate Professor (Grade IV)
of Computer Law and Legal Informatics at the
Faculty of Law of the University of the Republic - Uruguay
DOI: https://doi.org/10.59224/dti8.ch17

Abstract: Algorithmic constitutionalism redefines rights in the digital age. No longer mere limits on state power, rights become design principles embedded in algorithmic systems. Privacy demands data architectures that prevent invisible discrimination; freedom of expression must address moderation algorithms; equality requires representative training data. Guarantees shift from abstract rules to audits, transparency, and human oversight. Reconfiguration is central: dignity calls for explainability and autonomy, freedom evolves into cognitive freedom. Rights shift from the individual to the collective, ensuring justice and legitimacy in an algorithmic society.

Keywords: Algorithmic Constitutionalism – Fundamental Rights – Digital Governance

Table of contents: 1. Classical Constitutionalism and Its Historical Evolution; 2. Algorithmic Power as a Constitutional Challenge; 3. Algorithms as Instruments of Power; 4. Digital Governance and Algorithmic Constitutionalism; 5. Algorithmic Governance; 6. Ethical Principles for a Technological World; 7. From the Guarantee to the Reconfiguration of Fundamental Rights; 7.1. The transition from the individual to the collective; 7.2. Fundamental rights in the age of algorithms; 7.2.1. Privacy and data protection: from limit to design principle; 7.2.2. Equality and non-discrimi-

nation: towards algorithmic equity; 7.2.3. Freedom of expression and media diversity in digital environments; 7.2.4. Human dignity and autonomy in the algorithmic age; 8. Constitutional Culture in the Digital Age; 8.1. The hybrid nature of the contemporary world; 8.2. Modernising the Constitution; 8.3. Algorithmic constitutionalism and democratic legitimacy; 8.4. Algorithmic power and its implications for democracy; 8.5. Algorithmic constitutionalism as a paradigm of legitimacy; 9. Conclusions.

1. CLASSICAL CONSTITUTIONALISM AND ITS HISTORICAL EVOLUTION

Constitutionalism emerged to limit sovereign power and guarantee individual freedom against state arbitrariness. From the very first written constitutions, the aim was to establish a regulatory framework that would ensure the supremacy of the law, the separation of powers, and the protection of fundamental rights.

The Constitution stood as a political and juridical pact that simultaneously limited and legitimised the exercise of power. However, constitutionalism was never static: it transformed over time, adapting to new forms of power and new social demands.

In the 19th century, liberal constitutionalism focused on protecting individual freedoms against the State. In the 20th century, with the rise of social constitutionalism, fundamental rights expanded into the economic and social spheres, incorporating guarantees of material equality and distributive justice. Subsequently, multilevel and supranational constitutionalism introduced the dimension of universal human rights, linking national legal systems with international human rights protection systems.

Today, in the 21st century, we are facing a new development: algorithmic constitutionalism. This has emerged in response to a form of power that does not fit within the traditional frameworks of political or economic authority, but rather within technological mediation. Algorithms, as systems of automated decision-making, shape social, economic, and political life in a pervasive and silent manner. Algorithmic constitutionalism presents itself as the third major expansion of constitutionalism: from safeguards against state power to protection against private power, and now to a reconfiguration in the face of technological power.

2. ALGORITHMIC POWER AS A CONSTITUTIONAL CHALLENGE

Algorithmic power is characterised by its invisibility, its ubiquity, and its capacity to structure everyday life. It is not exercised through physical coercion or explicit legal rules, but through informational mediation. Algorithms decide what news we see, what products are offered to us, what educational or career paths are suggested to us, and what risks are assigned to us in credit or insurance systems. This power is diffuse, decentralised, and, in many cases, opaque.

Traditional constitutional theory was built on the premise that power was visible, centralised, and exercised by identifiable political institutions. Algorithmic power breaks with this logic. There is no single sovereign, but rather a multiplicity of technological actors who design, implement, and control automated decision-making systems. Democratic legitimacy is challenged when decisions affecting fundamental rights are made without public deliberation, transparency, or accountability.

Algorithmic constitutionalism thus emerges as a response to this new power. It aims to transfer the principles of constitutionalism—guarantee, limitation, control, and protection of human dignity—to the realm of digital governance. But it is not merely a matter of extending traditional safeguards: it is about reconfiguring fundamental rights in light of the new conditions of digital life.

Contemporary constitutionalism faces an unprecedented challenge: the emergence of algorithms as mediators of social, economic, and political life. Whilst constitutionalism originally arose as a response to the need to limit sovereign power and to guarantee fundamental rights against state arbitrariness, today it faces a diffuse, decentralised, and technologically sophisticated power that does not fit within traditional forms of political authority. Algorithms, as automated decision-making systems, not only condition the exercise of rights but also reshape their content, scope, and effectiveness. Hence, the notion of "algorithmic constitutionalism" is understood as the effort to translate the principles of constitutionalism—guarantee, limitation, control, and protection of human dignity—into the realm of digital governance.

3. ALGORITHMS AS INSTRUMENTS OF POWER

It is remarkable how a word that originated in the works of a medieval mathematician and astronomer has, over the years, become one of the cornerstones of computer science. Indeed, the name is linked to the Arab mathematician Al-Khwarizmi, who spent much of his career in Baghdad around the year 800 AD. There, he established a Centre for Advanced Scientific Research and devoted himself particularly to algebra and astronomy. His methods for solving equations and his treatise on numbers, translated into Latin as "Algoritmi de numero Indorum", have made him the earliest reference for the word "algorithm"[1].

An algorithm is a set of logical, ordered steps aimed at solving a problem, involving specific actions to process input data from an initial state and obtain a solution[2].

They are defined as sets of rules, provisions, standards, guidelines, or methodical instructions for problem-solving or determining results. They are often used in developing computer code, which is why their results are expressed in binary.

> *"The "generic" definition of an algorithm refers to an "ordered set of operations aimed at solving a problem". This definition extends beyond computer science and even the hard sciences. It conveys a couple of important conceptual attributes: the set of operations may not be unique, and solutions may vary in quality depending on different metrics"*[3].

They are sequences of instructions designed to solve problems or perform tasks. They are used to write code in a programming language and, in this way, obtain results in binary format.

---

1 Universidad de La Plata.- "¿Por qué pensar algoritmos es tan importante en informática?", in Revista Institucional de la Facultad de Informática N° 4, 2016. Available in: https://core.ac.uk/download/pdf/76496451.pdf (last access: 16/05/2026)

2 Fonden-Calzadilla, Juan Carlos; Stuart-Cárdenas, Mavis Lis; Rodríguez-Matos, Lianne.- "La algoritmización: requisito necesario para la solución de problemas con el empleo de un lenguaje de programación" in Luz: Educar desde la Ciencia, vol. 17, núm. 3, La Habana, 2018. Available in: https://www.redalyc.org/journal/5891/589167671004/589167671004.pdf (last access: 16/05/2026).

3 Williamson, Ben.- "Computing Brains: Learning Algorithms and Neurocomputation in the Smart City", in Information, Communication & Society, 2016, págs. 1 a 19.

In this century, algorithms form the basis of AI, executing instructions through various techniques to transform data into patterns of information, then into knowledge, and ultimately to automate tasks, make predictions, or issue forecasts.

However, in the current context, their function goes beyond mere instrumentalism. Algorithms have become instruments of power that organise social life.

Algorithmic subordination implies that human decisions—from the allocation of work tasks to the recommendation of cultural content—are mediated by systems operating under opaque logic, embedded in codes that are neither easily accessible nor open to scrutiny by those affected. This opacity creates a power imbalance between those who design and control the algorithms and those who are subject to them.

In the digital age, algorithms have ceased to be mere technical tools and have become, as has been said, genuine instruments of power, which means we must acknowledge their role in shaping subjectivities, organising work, and redefining citizenship. Their ability to organise, classify, and make decisions on fundamental aspects of social life places them at the centre of philosophical, sociological, and political debates.

In light of the above, it must be stated that algorithms are not neutral: they constitute structures that reflect economic, political, and cultural interests and produce new forms of connection.

4. DIGITAL GOVERNANCE AND ALGORITHMIC CONSTITUTIONALISM

Digital governance, as Manuel Castells[4] points out, is characterised by global interconnection, the decentralisation of actors, and technological mediation in decision-making.

"Capitalism itself has undergone a process of profound restructuring, characterised by

[4] Note that Manuel Castells does not explicitly use the term "digital governance" in his texts. That concept was later consolidated in academic literature. However, he does describe the "red society2, which provides the conceptual foundation for what is now understood as digital governance: global interconnection, decentralization of actors, and technological mediation.

greater flexibility in management; the decentralisation and interconnection of companies, both internally and in their relations with others… As a consequence of this general overhaul of the capitalist system… we have witnessed the global integration of financial markets… and the incorporation of valuable segments of economies worldwide into an interdependent system that functions as a single unit in real time""New information technologies are not merely tools to be applied, but processes to be developed. Users and creators can become the same. In this way, users can take control of technology, as in the case of the Internet"[5].

Algorithms are not mere technical instruments, but structures of power that determine what information circulates, what opportunities are offered, and what risks are assigned.

Shoshana Zuboff, in The Age of Surveillance Capitalism, warns that algorithmic logic can erode individual and collective autonomy, generating an invisible and ubiquitous power that eludes traditional mechanisms of democratic control.

"The threat looming over us is no longer that of a totalitarian "Big Brother" state, but that of an omnipresent digital architecture: a "Big Other" operating in the interests of surveillance capital"[6]*. "Surveillance capitalism unilaterally claims the human experience as free raw material to be translated into behavioural data…"*[7].

Francisco Balaguer Callejón[8] emphasises that constitutional democracy faces a radical challenge: algorithms can influence electoral processes, manipulate public opinion, and erode democratic legitimacy if they are not subject to constitutional oversight.

Algorithmic constitutionalism seeks to respond to this new reality. Fundamental rights can no longer be conceived solely as individual guarantees against the State,

---

5 Castells, Manuel.- La era de la información: Economía, sociedad y cultura. Vol. I: La sociedad red, Siglo XXI Editores, México, 1996, p. 58-59.

6 Zuboff, Shoshana.- La era del capitalismo de la vigilancia: La lucha por un futuro humano frente a las nuevas fronteras del poder, Ediciones Paidós, Barcelona, 2020, p. 24.

7 Zuboff, Shoshana.- La era del capitalismo de la vigilancia: La lucha por un futuro humano frente a las nuevas fronteras del poder, Ediciones Paidós, Barcelona, 2020, p. 94.

8 Balaguer Callejón, Francisco.- La constitución del algoritmo. Fundación Manuel Giménez Abad. Segunda Edición. Zaragoza, 2023.

but also as collective principles that guide the design and regulation of technological systems. Privacy is redefined as a digital common good, equality as algorithmic fairness, freedom of expression as the right to informational plurality, and human dignity as a requirement for autonomy in the face of cognitive manipulation. In all these cases, the guarantee becomes a reconfiguration: fundamental rights not only protect against violations, but also guide the architecture of digital governance.

Democratic legitimacy becomes the core of this reflection. Frank Pasquale notes that "*transactions that are too complex to be explained to outsiders may well be too complex to be allowed to exist*"[9] whilst noting that "*an intelligible society would ensure that the key decisions of its most important companies are fair, non-discriminatory and open to criticism*"[10], which allows us to assert that the opacity of algorithms erodes citizens' ability to understand and question the decisions that affect them.

Cathy O'Neil shows how algorithms can reproduce and amplify historical biases, creating structural inequalities that affect entire communities.

> *"The mathematical applications driving the data economy are based on decisions made by fallible human beings. Undoubtedly, some of these decisions were made with the best of intentions. However, many of these models incorporated human prejudices, misunderstandings, and biases into the software systems that increasingly managed our lives. Like gods, these mathematical models were opaque; their workings were invisible to all except the most eminent experts in their field: mathematicians and computer scientists. Their verdicts, even when erroneous or harmful, were indisputable and final. And they tended to punish the poor and oppressed in our society"*[11].

Luciano Floridi argues for an information ethics that protects human dignity. In this regard, concerning dignity as the core of information ethics:

> *"it assesses the duty of any rational being in terms of their contribution to the well-being and flourishing of the infosphere, and that any process, action, or event that negatively affects the entire infosphere—such as an increase in its level of metaphysical entropy—is*

---

9 Pasquale, Frank.- The Black Box Society: The Secret Algorithms That Control Money and Information. Harvard University Press, Cambridge, 2015, p. 23.

10 Pasquale, Frank.- The Black Box Society. Harvard University Press, Cambridge, 2015, p. 189.

11 O'Neil, Cathy.- Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy. Crown Publishing Group, Nueva York, 2016, p. 3.

an example of evil"[12],

directly linking information ethics with the protection of the dignity of informational subjects. Likewise, regarding the need for an ethics that protects people in the digital age, he states that:

> *"the emergence of the infosphere, as a new environment in which human beings spend a large part of their lives, explains the need to expand further the conception of what can be considered a moral patient"*[13],

thus justifying the creation of a specific ethics for the digital environment, centred on human dignity. He concludes that the moral value of an action depends on whether it contributes to the preservation, flourishing and dignity of informational entities,

> *"the more powerful humanity becomes as an agent, the greater are its duties and responsibilities to oversee not only the development of its own nature and habits, but also the well-being and flourishing of each of its ever-expanding spheres of influence, including the entire infosphere."*[14],

thereby emphasizing that dignity is the fundamental criterion for evaluating actions in the informational realm.

In short, all the authors cited agree that democratic legitimacy requires citizens to understand, question, and participate in the design of algorithmic systems.

Algorithmic constitutionalism, in this sense, is not merely an extension of classical constitutionalism but a conceptual transformation. Guarantees are transformed into reconfiguration, and reconfiguration becomes the core of protection. To be legitimate, digital governance must be constitutionalised: it must incorporate the values of dignity, equality, freedom, and participation into the very architecture of technological systems.

The relationship between algorithmic constitutionalism and digital governance underscores the need to rethink the linkages among technology, power, and rights. Algorithms become actors of power, democratic legitimacy is a structural require-

---

12 Floridi, Luciano.- The Ethics of Information. Oxford University Press, Oxford, 2013, p. 84.

13 Floridi, Luciano.- The Ethics of Information. Oxford University Press, Oxford, 2013, p. 112.

14 Floridi, Luciano.- The Ethics of Information. Oxford University Press, Oxford, 2013, p. 168.

ment of technological systems, and fundamental rights are design principles. This is the horizon of constitutionalism in the digital age: an algorithmic constitutionalism that guarantees democratic legitimacy and protects society against technological power.

5. ALGORITHMIC GOVERNANCE

Algorithmic governance constitutes one of the most sophisticated and urgent debates of the digital age. Algorithms, especially those based on artificial intelligence and machine learning, have ceased to be mere technical tools and have become structures of power that shape decisions, distribute opportunities, and condition freedoms.

Algorithmic governance provides the necessary mechanisms of control and transparency to ensure that digital systems respect fundamental rights and social values.

Thus, for example, the link between algorithmic governance and compliance is evident in the need to audit and supervise algorithmic systems that make decisions in sensitive areas such as criminal justice, employment, access to credit, or the distribution of social benefits. In all these cases, an algorithmic decision can directly affect people's lives, leading to exclusion or discrimination that is difficult to detect.

Compliance brings its expertise to bear here by building internal control frameworks and designing protocols that ensure the traceability of algorithmic decisions, the explainability of models, and the correction of biases.

Algorithmic governance, for its part, requires that these protocols be integrated into a broader system of digital social responsibility, in which transparency and public participation are central elements. The combination of both approaches enables progress towards a model of responsible innovation, capable of balancing technological efficiency with social justice.

Algorithmic opacity, often referred to as the "black box," is one of the main challenges. This is also "*conceived as a form of protection of intellectual property or corpo-*

rate secrecy, where intentional obscurity seeks to avoid scrutiny"[15].

Many artificial intelligence systems are so complex that not even their developers can clearly explain how they reach certain conclusions. This lack of explainability threatens the basic principles of responsibility and accountability, and, here, compliance becomes a strategic ally of algorithmic governance.

Through internal audits, codes of conduct, and oversight mechanisms, compliance can ensure that organisations do not merely deploy efficient algorithms, but also ensure their transparency and alignment with ethical values. Algorithmic governance, in this sense, is not merely a set of technical standards but becomes a democratic control system that requires the participation of multiple stakeholders: states, businesses, civil society, and international organizations.

The international dimension adds complexity to the relationship with algorithmic governance.

Algorithms operate in a transnational environment, where digital platforms and artificial intelligence systems are deployed beyond national borders. This creates tensions between local regulatory frameworks and global corporate practices, as well as the need to establish international standards that harmonize transparency and accountability criteria.

Algorithmic governance calls for the creation of universal principles of digital ethics applicable to diverse contexts. Regulatory fragmentation can create legal loopholes that companies exploit to operate in more lenient jurisdictions, reinforcing the urgency of multi-level governance capable of coordinating state, regional, and international efforts.

6. ETHICAL PRINCIPLES FOR A TECHNOLOGICAL WORLD

In light of the above and to address the ethical and political challenges of the technological world, it is deemed necessary to engage in broad deliberation within a

[15] Kirsten, Martin.- "Implicaciones éticas y responsabilidad de los algoritmos", in *Revista de Ética Empresarial*, 2019. Available in: https://www.researchgate.net/publication/324896361_Ethical_Implications_and_Accountability_of_Algorithms (last access: 16/05/2026).

framework of minimum principles and rules that different social groups can recognise as safeguarding their own interests and accept as the basis for a broad consensus.

I therefore believe that an ethical framework, comprising, in principle, four principles, would provide the context for reaching agreements on the techno-scientific controversies that confront us.

These principles are: responsibility, precaution, autonomy, and justice.

The proposed ethics is an interrelated system of these principles, in which the general rule is that they must be balanced and interwoven.

As principles, these ethical minimums should not be understood as rigid rules to be applied deductively, nor as decision-making algorithms.

These principles constitute a rational framework for making decisions by consensus, but they are neither automatic procedures nor procedures of absolute certainty; they do not replace, but rather depend on, a society's deliberative capacity to make more prudent, reasonable, and just decisions.

- The principle of responsibility determines what we are obliged to care for and preserve: the integrity and dignity of human life first and foremost; that is to say, the primary object of responsibility is a human life that retains its characteristics of ethical consciousness and free action, of finitude (birth and death), of vulnerability, of unity, and of biological-genetic community. The first moral obligation is, therefore, to ensure the existence of moral beings capable of responsibility.

Responsibility entails developing scientific knowledge to prevent and anticipate the negative effects of technological intervention, as well as disseminating, as widely as possible, what is known about the risks and harms. If the looming danger is shared (the global ecological crisis), the ethical response must be coordinated and the risks equitably shared. If the technological world is systemic and global, the ethical-political response must also be systemic and globalised.

- The precautionary principle, which states that, where there is a possibility of danger based on reasonable forecasts, even in the absence of conclusive scientific evidence, and if the potential harm is incalculable or outweighs the expected benefit, the planned technological action must be reviewed, halted, modified, or prevented.

Damage and risks must be kept at a rationally acceptable level, provided that they

do not entail an unfair distribution within society.

- The principle of autonomy must be interpreted to mean that every technological action must have the informed consent of the individuals and communities that will be directly benefited and, likely, affected. In the technological world, citizens must have the right to deliberate on techno-scientific innovations and to choose, with sufficient information, the risks they are willing to take in full knowledge of the facts, based on their own values and cultural parameters.

Technological actions must protect, promote, and enhance individual autonomy so that each person may make responsible decisions regarding their own body and life, without affecting or restricting the autonomy and freedom of others

- The principle of distributive justice regarding technological benefits, but also regarding risks. This means not only that more and more people benefit from technological development and that gaps are narrowing, but also that risks are reasonable and socially and internationally shared.

7. FROM THE GUARANTEE TO THE RECONFIGURATION OF FUNDAMENTAL RIGHTS

Algorithmic constitutionalism is not merely an extension of traditional constitutionalism to a new subject matter. It entails a conceptual transformation: fundamental rights are no longer conceived solely as limits on state power, but also as parameters for the design, implementation, and evaluation of algorithmic systems. In this sense, the guarantee becomes a reconfiguration. The right to privacy, for example, is not limited to preventing the state's unlawful intrusion into private life, but requires a rethinking of the architecture of data systems, machine learning models, and the inference logics that can produce invisible discrimination. The right to freedom of expression is redefined in the face of content moderation algorithms that determine which voices are amplified and which are silenced. The right to equality faces algorithmic biases that reproduce and amplify structural inequalities.

The first dimension of algorithmic constitutionalism is the guarantee. This is where the classical tradition lies: fundamental rights operate as normative limits against any form of power. However, the guarantee in the algorithmic sphere requires mechanisms that differ from traditional ones. Abstract rules alone are not

enough; technical audits, model transparency, traceability of automated decisions, and the capacity for human intervention are required. The guarantee becomes a dynamic process combining legal regulation, technological design, and institutional oversight.

The second dimension is that of reconfiguration. This is where the true conceptual innovation lies. Fundamental rights are not only protected against potential violations, but are also redesigned in light of the new conditions of digital life. Human dignity, the core of constitutionalism, translates into demands for algorithmic explainability, the prohibition of systems that manipulate behaviour subliminally, and the need to preserve spaces of autonomy in the face of hyper-personalisation of content. Equality is redefined as algorithmic equality, which requires that automated decision-making systems be trained with representative data and subjected to impact assessments. Freedom is transformed into cognitive freedom, linked to the protection of attention and the capacity for deliberation in the face of digital environments designed to capture and direct behaviour.

Algorithmic constitutionalism, therefore, is not merely an additional chapter in digital constitutionalism. It is an emerging paradigm that compels us to rethink the relationship between technology, power, and rights. Within this framework, fundamental rights become principles of regulatory and technical design. Safeguards shift towards the architecture of systems, and reconfiguration becomes the core of protection.

Traditional safeguards prove insufficient in the face of this power. Conventional oversight mechanisms—courts, parliaments, regulatory bodies—lack the technical capacity and speed required to intervene in real time within systems that process millions of data points per second. Hence, the need for reconfiguration: fundamental rights must be translated into technical requirements embedded in the design from the outset. The notion of "privacy by design" or "ethics by design" is a manifestation of this shift. Algorithmic constitutionalism demands that rights be not merely external limits, but also internal parameters for technological construction.

Algorithmic constitutionalism, in short, opens up a new horizon for constitutional theory. Fundamental rights are no longer merely safeguards against power, but design principles in the face of technology. The guarantee becomes reconfiguration,

and reconfiguration becomes the contemporary form of protection.

In short, algorithmic constitutionalism represents a profound transformation of contemporary constitutionalism. We move from the classical guarantee against state power to the reconfiguration of fundamental rights in the face of technological power. As indicated, rights become design principles, the guarantee becomes a technical-legal process, and reconfiguration becomes the core of protection. This paradigm not only redefines constitutional theory but also opens up a new horizon for democratic practice and the protection of human dignity in the digital age.

7.1. The transition from the individual to the collective

A constant tension between the individual and collective dimensions has marked the history of fundamental rights. From the earliest formulations of classical liberalism, fundamental rights were conceived as individual safeguards against state power: freedom of expression, privacy, property, and equality before the law. These rights were intended to protect the individual against the sovereign's arbitrariness. However, over time, the evolution of constitutionalism and modern societies has shown that rights cannot be reduced to a strictly individual dimension. Social justice, material equality, and the protection of the commons demanded a broader conception that incorporated the collective dimension.

From Locke and Montesquieu to the 18th-century declarations of rights, the central idea was to protect the individual against state arbitrariness. Freedom, property, privacy, and equality were understood as personal attributes, linked to the autonomy and dignity of every human being. However, as Norberto Bobbio[16] pointed out, rights are not static: they evolve and respond to new social and technological conditions.

Today, in the 21st century, the emergence of artificial intelligence and algorithmic systems poses a new challenge. Fundamental rights, traditionally conceived as individual rights, must be reconfigured as collective rights. The reason is clear: algorithms operate on vast volumes of data, affect entire communities, and generate structural

[16] Bobbio, Norberto.- A era dos direitos. Trad. De Carlos Nelson Coutinho, Elsevier, Rio de Janeiro, 2004.

impacts that transcend the individual sphere. Privacy, equality, freedom of expression, and human dignity can no longer be protected solely at the individual level; they require collective protection that ensures algorithmic justice and democratic legitimacy.

This transition from the individual to the collective is neither linear nor uniform. It is a historical and conceptual process that reflects transformations in society, the economy, and technology.

In the 19th century, liberal constitutionalism focused primarily on protecting individual freedoms. In the 20th century, social constitutionalism incorporated collective rights in the areas of work, education, and health. Today, in the 21st century, the emergence of artificial intelligence and algorithmic systems poses a new challenge: fundamental rights, traditionally conceived as individual rights, must be reconfigured as collective rights.

Algorithmic systems operate on vast volumes of data, affect entire communities, and generate structural impacts that transcend the individual sphere.

7.2. Fundamental rights in the age of algorithms

Contemporary constitutional theory faces an unprecedented challenge: the emergence of algorithms as mediators of social, economic, and political life. Whilst fundamental rights were originally conceived as limits on state power. As safeguards against arbitrariness, today they are confronted with a diffuse, decentralised, and technologically sophisticated power that does not fit within the classical forms of political authority. Hence, the need to consider fundamental rights through an algorithmic lens, that is, as principles that not only protect against external infringements but also guide the design, implementation, and evaluation of technological systems that directly affect people's lives.

This algorithmic shift implies a conceptual transformation. Fundamental rights are no longer conceived solely as safeguards against political or economic power, but also as normative and technical parameters that must be integrated into the architecture of digital systems. Privacy, equality, freedom of expression, and human dignity are redefined in light of the new conditions of digital life. Protection becomes recon-

figuration, and the safeguard shifts towards technological design.

7.2.1. Privacy and data protection: from limit to design principle

The right to privacy and the protection of personal data are among the areas most affected by algorithmic logic. In classical constitutionalism, privacy was conceived as a limit on the State's illegitimate intrusion into private life. In algorithmic constitutionalism, privacy becomes a principle of data governance that demands transparency in the processes of data collection, anonymisation, storage, and use.

Algorithms operate on large volumes of data, generating profiles and predictions that can be invasive and discriminatory. Privacy, then, is not merely about preventing unauthorised access, but requires a rethinking of the architecture of data systems. The notion of "privacy by design" reflects this transformation: the right becomes a parameter of technological design. The guarantee is reconfigured, and data protection becomes a structural principle of algorithmic systems.

Algorithmic privacy also entails protection against non-consensual inferences. Artificial intelligence systems can deduce sensitive information from seemingly neutral data. Privacy, therefore, demands limits on inference capacity and mechanisms to control the use of profiles. Transparency and explainability become constitutional requirements: citizens must be able to understand how their data is used and the consequences of automated decisions.

7.2.2. Equality and non-discrimination: towards algorithmic equity

The right to equality faces a particular challenge in the algorithmic sphere. In classical constitutionalism, equality was protected against explicit discrimination. In the digital sphere, discrimination can be implicit, statistical, or emerge from data patterns. Artificial intelligence systems can reproduce and amplify historical biases, generating invisible inequalities.

Equality, then, is reconfigured as a principle of algorithmic fairness. This requires that automated decision-making systems be trained with representative data, subjected to impact assessments, and audited to detect biases. Equality is no longer merely an external constraint, but also an internal design parameter.

Algorithmic fairness also implies the need for corrective mechanisms. Systems must be able to identify and correct biases in real time. Equality becomes a dynamic requirement, linked to the system's ability to adapt and ensure fair outcomes. Transparency regarding decision-making criteria and the possibility of challenging automated decisions form part of this reconfiguration.

7.2.3. Freedom of expression and media diversity in digital environments

Freedom of expression is radically transformed in the algorithmic sphere. In classical constitutionalism, it was protected against state censorship. In the digital sphere, censorship can be indirect, exercised through the invisibilisation of content by recommendation systems. Freedom of expression is reconfigured as a right to fair visibility, to media diversity, and to freedom from algorithmic manipulation. This guarantee becomes a demand for transparency regarding moderation criteria and the possibility of challenging automated decisions. Freedom is no longer merely the absence of censorship, but also the ability to participate in a pluralistic and equitable digital public sphere.

The reconfiguration of freedom of expression also implies protection against cognitive manipulation. Recommendation systems can direct users' attention to certain content, thereby affecting their capacity for deliberation. Freedom is redefined as cognitive freedom, linked to the protection of attention and autonomy against digital environments designed to capture and direct behaviour.

7.2.4. Human dignity and autonomy in the algorithmic age

Human dignity, the cornerstone of constitutionalism, faces a profound challenge in the realm of algorithms. Artificial intelligence systems can manipulate behaviour subliminally, undermine people's autonomy, and diminish their decision-making capacity.

Algorithmic dignity entails protection against the hyper-personalisation of content. Digital systems can generate information bubbles that limit the diversity of perspectives and reduce the capacity for deliberation. Dignity demands the preserva-

tion of spaces of plurality and autonomy, ensuring that citizens can access diverse information and make free decisions.

The reconfiguration of human dignity is also linked to the need for democratic participation in the design of algorithmic systems. The legitimacy of automated decisions depends on citizens' ability to understand them, question them, and participate in their development. Dignity becomes a principle of algorithmic participation, requiring mechanisms for public deliberation on the use of artificial intelligence.

8. CONSTITUTIONAL CULTURE IN THE DIGITAL AGE

Constitutional culture, understood as the set of values, principles, and narratives that underpin the Constitution's relevance in a society's political and legal life, is currently experiencing a crisis of legitimacy and relevance.

It is worth noting that this culture is currently undergoing a regression, driven by economic and technological transformations within the framework of the so-called "*third globalisation*"[17]. In this context, technological and digital advances appear to offer immediate and satisfactory responses to social needs, whilst the Constitution is seen as a constraint: a regulatory requirement that threatens to stifle the dynamism of major technology companies.

Consequently, constitutional culture must reinvent itself to become a productive force that enables the reconciliation of fundamental rights with new technological landscapes.

This third wave of globalisation, characterised by digital interconnection, the expansion of technological platforms, and the hegemony of the data economy, has radically transformed social, political, and economic interactions. In this scenario, the values associated with economic and technological progress have supplanted traditional constitutional values, such as sovereignty, the separation of powers, and the supremacy of fundamental rights.

---

[17] Huang, Guanxiang.- "The Three Waves of Globalization: Evolution, Impacts, and Future Challenges", in *Highlights in Business, Economics and Management*, Volumen 54, 2025. Available in: file:///C:/Users/User/Downloads/The_Three_Waves_of_Globalization_Evolution_Impacts.pdf (last access: 16/05/2026).

The Constitution, historically conceived as the supreme legal framework for political life, now appears to be an inadequate instrument for regulating phenomena that transcend national borders. Large technology companies operate in a global arena governed by private law, whilst state constitutions are limited to regulating the physical reality within their borders. This asymmetry creates a perception that constitutional culture is obsolete, overshadowed by the immediacy and effectiveness of technological solutions.

8.1. The hybrid nature of the contemporary world

However, today's world is characterised by a hybrid configuration, marked by fragmentation and the clash between two realities:

- Digital reality: governed by private law, with a global reach and dominated by technology corporations that design algorithms and platforms according to their business models.
- Physical reality: governed by public law, with a state-level scope, where the analogue Constitution remains the fundamental regulatory framework.

Added to this division is the tension between the analogue Constitution, which regulates the physical world, and the digital Constitution, whose regulation of virtual reality and artificial intelligence is still in its infancy.

> *"The problem is not merely that the constitution addresses a facet of reality that has all but ceased to exist (the analogue world), but that it fails to regulate the reality that has taken hold and is shaping a new kind of society living in a digital world"*[18].

The lack of a comprehensive constitutional framework for the digital world creates regulatory gaps that technology companies exploit to consolidate their power.

Against this backdrop, constitutional culture cannot limit itself to passively resisting the onslaught of technology. It must offer new narratives and doctrinal constructs that re-evaluate its position within the legal system and in the public sphere. The

[18] Balaguer Callejón, Francisco.- La Constitución del Algoritmo, Fundación Manuel Giménez Abad, Available in: https://www.fundacionmgimenezabad.es/sites/default/files/Publicar/actividades/documentos/2021/20210415_dc_resumen_ponencia_balaguer_callejon_f_es_o.pdf (last access: 16/05/2026).

Constitution must be updated to address the challenges of the digital society, leveraging its regulatory capacity grounded in fundamental rights.

The tension between the digital and physical worlds should not be seen as a threat, but as an opportunity to drive new civilizational developments. Constitutional culture can become a productive force if it articulates the principles of freedom, equality, and human dignity in relation to technological dynamics, ensuring that digital progress does not come at the expense of democracy and fundamental rights.

8.2. Modernising the Constitution

Modernising the Constitution involves adapting it to new social processes and aligning it with the realities of the digital society. However, this process must be carried out with full awareness of the limitations of a national constitution. The problems arising from technological globalisation cannot be resolved solely at the national level, as they require global solutions.

The Constitution's task is, therefore, to distinguish between the global causes of problems and their domestic effects, to determine the possible spheres of intervention for public law. This entails recognising that, although technology companies operate globally, their impacts are felt within national societies, where the Constitution remains an indispensable normative benchmark.

A fundamental aspect that must not be overlooked is that the Constitution embodies a civilizational legacy that must be passed on to future generations. Fundamental rights, the democratic organisation of power, and the guarantee of human dignity are historic achievements that cannot be sacrificed in the name of technological progress.

Constitutional reflection must distinguish between the structural characteristics of technological development and the configurations arising from business models. Only in this way can it inspire a future digital constitutional process that preserves democracy's essential values and extends them into the virtual world.

In this regard, one of the most pressing challenges facing legal scholarship is to resolve the tension between the digital world and democracy, as well as between fundamental rights and the Constitution itself.

Digitising the Constitution is not enough; it is also necessary to constitutionalise algorithms and virtual reality.

This entails subjecting automated decisions and the dynamics of digital platforms to constitutional principles, ensuring that they respect equality, privacy, freedom of expression, and other fundamental rights. The Constitution cannot remain detached from digital reality; it must undergo a profound transformation to incorporate it into this new sphere fully.

> *"An algorithm cannot serve as a license for the mass infringement of rights. The Constitution cannot continue to turn its back on a reality such as the digital world, in which a large part of economic activity and people's lives now take place"*[19].

It is not simply a matter of adding new provisions to the Constitution to regulate algorithms and virtual reality. What is at stake is a structural transformation of the Constitution itself, which must take into account the profound changes that the digital reality entails for its nature and function.

8.3. Algorithmic constitutionalism and democratic legitimacy

Algorithmic constitutionalism not only redefines fundamental rights but also opens up new horizons for democratic legitimacy. By making decisions that affect rights, algorithms become actors of power. However, they lack political legitimacy. This must ensure that automated decisions are subject to democratic control, that citizens have the capacity to understand and question the systems, and that there is transparency and accountability.

Democratic safeguards are being reconfigured as algorithmic participation, which requires mechanisms for public deliberation on the design and use of artificial intelligence systems. Democracy cannot be limited to controlling political institutions; it

---

[19] Balaguer Callejón, Francisco.- La Constitución del Algoritmo, Fundación Manuel Giménez Abad, Available in: https://www.fundacionmgimenezabad.es/sites/default/files/Publicar/actividades/documentos/2021/20210415_dc_resumen_ponencia_balaguer_callejon_f_es_o.pdf (last access: 16/05/2026).

must extend to the technological systems that impact everyday life.

In conclusion, algorithmic constitutionalism represents a profound transformation of contemporary constitutionalism. We are moving from the traditional safeguard against state power to a reconfiguration of fundamental rights in the face of technological power. The relationship between algorithmic constitutionalism and democratic legitimacy is therefore central to understanding the future of constitutional theory.

8.4. Algorithmic power and its implications for democracy

Identifiable political institutions do not exercise algorithmic power but are mediated by technological systems designed by both public and private actors. This power is diffuse and decentralised, yet profoundly structural. Algorithms determine which voices are amplified in the digital public sphere, which citizens are afforded certain opportunities, and which risks are assigned to them in credit or insurance systems.

Democracy, in its classical conception, is based on the idea that political power must be legitimate, that is, subject to the will of the people and to mechanisms of institutional control. Algorithmic power challenges this logic. Decisions affecting fundamental rights are taken without public deliberation, transparency, or accountability. Democratic legitimacy is eroded, as citizens lack the capacity to understand and question the systems that shape their daily lives.

Algorithmic constitutionalism has emerged as a response to this challenge. It aims to apply the principles of constitutionalism—guarantee, limitation, control, and the protection of human dignity—to the realm of digital governance. But it is not merely a matter of extending traditional safeguards: it is about reconfiguring fundamental rights in light of the new conditions of digital life.

8.5. Algorithmic constitutionalism as a paradigm of legitimacy

Algorithmic constitutionalism proposes a new paradigm of democratic legitimacy. In classical constitutionalism, legitimacy was linked to the will of the people and to the institutional control of political power. In the algorithmic sphere, legitimacy is redefined as the ability of citizens to participate in the design and control of the

technological systems that affect their rights.

Algorithmic legitimacy demands transparency, explainability, and accountability. Systems must be comprehensible to citizens, and automated decisions must be open to scrutiny. Democratic participation becomes a structural requirement: citizens must have the capacity to deliberate on the use of artificial intelligence and to influence its design.

Algorithmic constitutionalism, in this sense, is not merely an extension of classical constitutionalism but a conceptual transformation. The guarantee is transformed into reconfiguration, and reconfiguration becomes the core of protection.

9. CONCLUSIONS

Constitutional culture today faces a historic challenge: to reinvent itself in the context of the third wave of globalisation and the digital society. The perception of obsolescence afflicting it must not lead to its abandonment, but rather to its updating. Digitising the Constitution and constitutionalising algorithms are urgent tasks that will enable fundamental rights to be aligned with technological dynamics, preserving the legacy of civilisation and projecting it into the future.

The Constitution must cease to be seen as a constraint that hinders innovation and become a framework that guides technological development towards ends compatible with human dignity and social justice. This requires a new constitutional culture, capable of engaging with technology without renouncing its fundamental principles.

Fundamental rights in an algorithmic context represent a profound transformation of contemporary constitutionalism. We are moving from the traditional safeguard against state power to a reconfiguration in the face of technological power.

Fundamental rights are no longer merely external limits, but also internal parameters of technological construction. Privacy, equality, freedom, and dignity become structural requirements of algorithmic systems.

Ultimately, this new constitutionalism calls for a rethink of the relationship among technology, power, and rights. Safeguards are transformed into reconfigurations, and reconfigurations become the contemporary form of protection.

This is the horizon of constitutionalism in the digital age: an algorithmic constitutionalism that safeguards human dignity amid technological power.

The emergence of this new constitutional paradigm—now algorithmic—represents a necessary response to the growing imbalance between global technological power and state regulatory structures. It is not merely a matter of incorporating regulatory provisions into analogue constitutions, but of rethinking the constitutional architecture from the ground up, integrating algorithms and artificial intelligence into the framework of fundamental rights. This new paradigm must ensure that the programming, design, and application of digital systems are guided by the principles of dignity, freedom, and equality, thereby preventing the market's instrumental logic from eroding the civilising values embodied in the Constitution. In this sense, algorithmic constitutionalism is not a mere technical adjustment, but a cultural and legal transformation that seeks to restore the centrality of the human being in a world dominated by automation and digital interconnection.

The task of safeguarding human dignity in the face of technological power requires the Constitution to become a framework for resistance and guidance, capable of harnessing the power of algorithms and global platforms. The digitisation of the Constitution must be accompanied by a process of constitutionalising virtual reality, in which fundamental rights serve as limits and guides for technological development. Only in this way will it be possible to prevent artificial intelligence and the data economy from becoming dehumanising forces that subjugate social life to logics of control and dependency. Algorithmic constitutionalism, in short, is envisaged as the instrument that will reconcile technological innovation with democracy, ensuring that digital progress unfolds in harmony with human dignity and with the civilizational legacy that constitutions have preserved throughout history.

REFERENCES

Balaguer Callejón, Francisco. La constitución del algoritmo. Fundación Manuel Giménez Abad, 2ª ed., Zaragoza, 2023. Available at: https://www.fundacionmgimenezabad.es/sites/default/files/Publicar/actividades/documentos/2021/20210415_dc_resumen_ponencia_balaguer_callejon_f_es_o.pdf (last access: 16 May 2026).

Bobbio, Norberto. A era dos direitos. Trad. Carlos Nelson Coutinho. Elsevier, Rio de Janeiro, 2004.

Castells, Manuel. La era de la información: Economía, sociedad y cultura. Vol. I: La sociedad red. Siglo

XXI Editores, México, 1996.

Floridi, Luciano. The Ethics of Information. Oxford University Press, Oxford, 2013.

Fonden-Calzadilla, Juan Carlos; Stuart-Cárdenas, Mavis Lis; Rodríguez-Matos, Lianne. "La algoritmización: requisito necesario para la solución de problemas con el empleo de un lenguaje de programación", Luz: Educar desde la Ciencia, vol. 17, núm. 3, La Habana, 2018. Available at: https://www.redalyc.org/journal/5891/589167671004/589167671004.pdf (last access: 16 May 2026).

Huang, Guanxiang. "The Three Waves of Globalization: Evolution, Impacts, and Future Challenges", Highlights in Business, Economics and Management, vol. 54, 2025.

Kirsten, Martin. "Implicaciones éticas y responsabilidad de los algoritmos", Revista de Ética Empresarial, 2019. Available at: https://www.researchgate.net/publication/324896361_Ethical_Implications_and_Accountability_of_Algorithms (last access: 16 May 2026).

O'Neil, Cathy. Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy. Crown Publishing Group, Nueva York, 2016.

Pasquale, Frank. The Black Box Society: The Secret Algorithms That Control Money and Information. Harvard University Press, Cambridge, 2015.

Universidad de La Plata. "¿Por qué pensar algoritmos es tan importante en informática?", Revista Institucional de la Facultad de Informática N° 4, 2016. Available at: https://core.ac.uk/download/pdf/76496451.pdf (last access: 16 May 2026).

Williamson, Ben. "Computing Brains: Learning Algorithms and Neurocomputation in the Smart City", Information, Communication & Society, 2016, p. 1-19.

Zuboff, Shoshana. La era del capitalismo de la vigilancia: La lucha por un futuro humano frente a las nuevas fronteras del poder. Ediciones Paidós, Barcelona, 2020.

REGULATING DEEPFAKES IN BRAZIL: INSIGHTS FROM INTERNATIONAL EXPERIENCE

Helena Penna Marchi

Master of Laws (UFMG). Lawyer.

DOI: https://doi.org/10.59224/dti8.ch18

Abstract: This article addresses the evolution of deepfake regulation, aiming to present the foundations for regulating the subject matter in Brazil in light of international experience. Initially, it discusses the definition of deepfakes and their characteristics, then explores the potential harms caused by deepfakes. Subsequently, the paper analyzes the different regulatory approaches adopted in three global contexts: China, the European Union, and the United States. Finally, the article discusses the pathways for regulating deepfakes in Brazil, where there is still no comprehensive specific legislation.

Keywords: deepfakes, regulation

Table of contents: 1. Introduction; 2. Deepfakes: definition and characteristics; 3. Potential harms caused by deepfakes and grounds for regulation; 4. Deepfakes regulation; 4.1. Deepfakes regulation in China; 4.2. Deepfakes regulation in the European Union; 4.3. Deepfakes regulation in the United States; 4.4. Summary of regulatory models; 5. Paths for Regulations of the matter in Brazil; 6. Conclusion; References.

1. INTRODUCTION

The idea of machines capable of autonomously replicating and understanding human reasoning was, until a few years ago, merely an imaginary construct reinforced by Hollywood's movies. Artificial intelligence (AI) was seen as an intangible

project, far from realization.

However, one of the intrinsic characteristics of new technologies is their dynamism and the speed with which they evolve, enabling increasing modernization and facilitating user access. What once seemed intangible is now part of everyday life, and often AI systems go unnoticed by their users (examples include assistant services on mobile phones, autonomous vehicles, and ChatGPT).

With this constant evolution and innovation comes the need for ongoing regulation of their effects, and countries around the world face the challenging task of creating effective regulations that can evolve with the same dynamism.[1] Despite consensus on the need for regulation, researchers, governments, and corporations present various proposals, and currently, there is no single regulatory standard.

According to Finocchiaro,[2] there are two approaches to AI regulation: the first consists in regulating AI as a whole, as done in the European Union and partially in Brazil; the second is based on specific regulations for AI applications in particular sectors, as practiced in the United Kingdom and China. Neither model is free from criticism, but their coexistence enables the study of the most appropriate form of regulation, facilitating the development of relevant principles and norms within the AI domain.

Regardless of the adopted model, it is necessary to create a system, or adapt an existing one,[3] that broadly and continuously encourages safe and responsible practices,[4] including enabling the coexistence of various national regulatory models. The combination of a joint effort to develop adequate algorithmic solutions and efficient institutional regulation is essential, with regulation needing to address technical aspects as well as social dimensions to adequately respond to the complexity of the

---

1 SOLUM, Lawrence B. Legal Personhood for Artificial Intelligences. North Carolina Law Review. Chapel Hill: The University of North Carolina School of Law, vol. 70, no. 04, Apr. 1992, p. 1231.

2 FINNOCCHIARO, Giusella. The Regulation of Artificial Intelligence. AI & Society. New York: Springer Nature, vol. 39, pp. 1961–1968, April 2023, pp. 1963–1964.

3 NARAYANAN, Arvind; KAPOOR, Sayash. AI Snake Oil: What Artificial Intelligence Can Do, What it Can't, and How to Tell the Difference. Princeton: Princeton University Press, 2024, p. 273.

4 ARBEL, Yonathan; MATTHEW, Tokson; LIN, Albert. Systemic Regulation of Artificial Intelligence. Arizona State Law Journal. Phoenix: Arizona State University, vol. 56, no. 2, pp. 545–620, Summer 2024, p. 618.

involved risks.[5]

In this context of growing attention to AI regulation globally, this study focuses on the emergence in recent years of so-called deepfakes,[6] which are nothing more than the result of manipulation, through artificial intelligence, of videos, audios, sounds, and other media, enabling the creation of new content that is different from the original, fictitious, and extremely realistic.[7]

The analysis proposed here has become increasingly necessary due to the growing ease of access to tools for creating deepfakes, allowing the production of realistic synthetic media by anyone with internet access. For example, in May 2025, Google launched Veo3,[8] an AI system that transforms textual descriptions into videos with automatically generated scenarios, characters, and speech, facilitating deepfake creation and in November of the same year the "Nano Banana Pro" application has been integrated into Google's Gemini generative AI app, which is capable of creating realistic images based on user prompts.[9]

---

5 CAJUEIRO, Daniel O.; CELESTINO, Victor R. R. A Comprehensive Review of Artificial Intelligence Regulation: Weighing Ethical Principles and Innovation. Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4950727, Access: May 10, 2026.

6 STRAUB, Sebastian. Der neue Rechtsrahmen für Künstliche Intelligenz – Ein Überblick zur geplanten EU Regulierung von KI. Available at: https://www.wolterskluwer.com/de-de/expert-insights/der-neue-rechtsrahmen-fuer-kuenstliche-intelligenz, Access: May 10, 2026.

7 "*Deepfake, in sum, is a technique of image or sound synthesis by means of AI. Its use enables the substitution of one person by another, modification of speech content, among numerous editing alternatives. Although usually associated with video production, nothing prevents its application in image or audio files alone. In practice, the term is also used to identify the video, audio, or image resulting from manipulation*." SCHREIBER, Anderson; RIBAS, Felipe; MANSUR, Rafael. *Deepfakes: regulação e responsabilidade civil*. In: TEPEDINO, Gustavo; SILVA, Rodrigo da Guia (Coord.). O direito civil na era da inteligência artificial. São Paulo: Thomson Reuters Brasil, 2020. p. 611.

8 Available at: https://www.cnnbrasil.com.br/tecnologia/saiba-como-criar-videos-gerados-por-ia-com-o-veo-3-do-google/, Access: May 10, 2026.

9 Available at: <https://gemini.google.com/app?utm_medium=paid-media&utm_medium=cpc&utm_campaign=nano_banana_pro_sem_bkws&utm_campaign=2024ptBR_gem-feb&utm_source=google&gad_source=1&gad_campaignid=22980692842&gbraid=0AAAAApk5Bhn1MLn_FRn1TCcmYBzWb5I3r&gclid=CjwKCAiA3fnJBhAgEiwAyqmY5WvGkgy7kC0m8gB1aM74WqO9XAmxyQJccnOCGBufQUuKIet-

Falling within the spectrum of generative artificial intelligence, it is expected that a combination of technology and norms will suffice to address the challenges arising from the creation and distribution of deepfakes.

In this context, aiming to analyze the best regulatory approach that balances technological advancement and social protection, this article will initially analyze what deepfakes are and the grounds for their regulation; secondly, it will examine existing regulations on the subject in the European Union, China, and the United States. Finally, in a third part, it will analyze, in light of existing models, the ideal form of regulation for this matter in Brazil.

2. DEEPFAKES: DEFINITION AND CHARACTERISTICS

The term "deepfake" is composed of distinct concepts, being a neologism formed by the union of the English words "fake" and "deep." The first is understood by its common meaning, "false." The second refers to the term "deep learning," which, in the field of artificial intelligence, constitutes a subcategory of machine learning, embodied in a learning technique by which computers learn to perform tasks that are natural for the human brain.[10]

According to Chesney and Citron's definition, deepfakes "*leverages machine-learning algorithms to insert faces and voices into video and audio recordings of actual people and enables the creation of realistic impersonations out of digital whole cloth*".[11] The result of such manipulation are realistic-looking media that make it look like someone said or did something they didn't.

Based on this concept, Medon highlights two characteristics of deepfakes: "*the use of advanced computational techniques, commonly artificial intelligence, as well as the*

fR2-yVBoCSDgQAvD_BwE&gclsrc=aw.ds>, Access: May 10, 2026.

10 "*Deep learning is a subset of machine learning in which the tasks are broken down and distributed onto machine learning algorithms that are organized in consecutive layers. Each layer builds upon the output from the previous layer. Together the layers constitute an artificial neural network that mimics the distributed approach to problem-solving carried out by neurons in a human brain*." Available at: http://webfoundation.org/docs/2017/07/AI_Report_WF.pdf, Access: May 10, 2026.

11 CHESNEY, Bobby; CITRON, Daniele. Deepfakes: a looming crisis for national security, democracy and privacy. *107 California Law Review*, 2019, p. 1758.

very high degree of realism that makes it almost impossible to detect the fraud."[12] Regarding the topic and in light of the distinction between strong AI systems and weak AI systems,[13] the Brazilian authors Schreiber, Ribas, and Mansur,[14] in agreement with Ramatik's understanding,[15] clarify that given the low degree of AI autonomy in the creation process, deepfakes would be classified as weak AI systems.

Furthermore, deepfakes are characterized as a form of generative artificial intelligence due to their capacity to autonomously produce new content from pre-existing models, rather than merely replicating or processing them. From a technical standpoint, two key elements are the variational autoencoders and the generative adversarial networks (GANs), which are used to learn complex patterns from large volumes of real data.[16] Once trained, these algorithms can generate new samples that mimic the characteristics of the original data, such as facial expressions, lip movements, and voice inflections, rendering them indistinguishable from authentic productions to the human eye.[17]

It is precisely this impossibility of naked-eye distinction that makes the study and regulation of deepfakes relevant, since with these innovative techniques it is possible

---

12 MEDON, Filipe. O direito à imagem na era das *deepfakes*. *Revista Brasileira de Direito Civil – RBD Civil*. Belo Horizonte, v. 27, p. 251-277, jan./mar. 2021, p. 263.

13 SEARLE, John. *Mente Cérebro e Ciência*. Translated by Artur Morão. Lisbon: Biblioteca de Filosofia Contemporânea, 2000, p. 36; BOSTROM, Nick. *Superinteligência. Caminhos, Perigos e Estratégias para um novo mundo*. Translated by Aurélio Antônio Monteiro et al. Rio de Janeiro: DarkSide, 2018, p. 38.

14 SCHREIBER, Anderson; RIBAS, Felipe; MANSUR, Rafael. Deepfakes: regulação e responsabilidade civil. In: TEPEDINO, Gustavo; SILVA, Rodrigo da Guia (Coord.). *O direito civil na era da inteligência artificial*. São Paulo: Thomson Reuters Brasil, 2020, p. 611.

15 RAMATIK, Amela. *Küsntliche Inteligenz im Rahmen der EU*. Johanes Kepler Universität Linz, 2023, p. 23. Available at: https://epub.jku.at/obvulihs/content/titleinfo/8631729/full.pdf, Access: May 10, 2026.

16 MITRA, Alakananda; MOHANTY, Saraju; KOUGIANOS, Elias. The World of Generative AI: Deepfakes and Language Models. Available at: https://doi.org/10.48550/arXiv.2402.04373, Access: May 10, 2026.

17 SHOAIB, Mohamed; WANG, Zefan; AHVANOOEY, Milada; ZHAO, Jun. Deepfakes, Misinformation, and Disinformation in the Era of Frontier AI, Generative AI, and Large AI Models. International Conference on Computer and Applications (ICCA), 2023, p. 2.

to manipulate authentic creations and images to give rise to new media.[18] As a tool for generating simulated reality, it involves risks ranging from defamation and fraud to mass disinformation, so its regulation demands approaches that consider not only the harm caused but also the artificial and intentional origin of the content.

In a document entitled Tackling Deepfakes in European Policy, the European Parliamentary Research Service highlighted five distinct timeframes linked to the deepfake lifecycle where regulation can be applied: (i) the technology itself; (ii) the creation of a deepfake; (iii) the circulation of such content; (iv) the target audience (if any) of this content; and (v) the audience accessing the content.[19] Essentially, the regulatory question is at which point in the deepfake lifecycle policymakers should attempt to regulate deepfakes.

As the primary characteristic of such creations is the distortion of reality, it is common for deepfakes to be automatically associated with an illicit act, embodied in false media, with the intent to harm the person portrayed or third parties. In this regard, some examples stand out, such as images of the arrest of former American President Donald Trump,[20] or the video in which former American President Barack Obama calls his then-rival Donald Trump "a complete idiot."[21] In Brazil, especially during the 2022 electoral period, numerous cases of deepfakes involving the main candidates for the Presidency of the Republic were reported.[22]

---

[18] "*Indeed, with these ultra-innovative techniques, it is possible to manipulate authentic audio, images, and videos to create false and distorted multimedia content: just a few images and an internet connection are enough to paste one person's face onto another's body and make them say and do things they never actually said or did.*" CAZZANIGA, Martina. A new technique (also) to convey disinformation: the European responses to deepfakes. In: Rivista di Diritto dei Media, vol. 1/2023, p. 172.

[19] Available at: https://www.europarl.europa.eu/RegData/etudes/STUD/2021/690039/EPRS_STU(2021)690039_EN.pdf, Access: May 12, 2026.

[20] Available at https://www.washingtonpost.com/politics/2023/03/22/trump-arrest-deepfakes/, Access: May 12, 2026.

[21] Available at https://www.businessinsider.com/obama-deepfake-video-insulting-trump-2018-4, Access: May 12, 2026.

[22] For example, a video in which journalist Renata Vasconcellos announces a false poll result (available at https://www.youtube.com/watch?v=X3tU01FDPhA, accessed July 9, 2023), as well as the Supreme Federal Court's order to remove a deepfake showing then-presidential candidate

In 2024, a video circulated on Instagram and TikTok featuring Elon Musk promoting an investment in crypto-assets. The video led several individuals to make considerable investments in the promoted company, which ultimately was revealed to be a fraudulent scheme through which the agents appropriated the invested funds.[23]

In the same vein of illegality, the German magazine *Die Aktuelle* published an interview attributed to the pilot Michael Schumacher, whose health condition has remained unknown since the accident he suffered in 2013.[24] The answers were generated by artificial intelligence (deepfake) without the family's authorization, who subsequently filed a lawsuit against the magazine.[25]

There is also a series of examples involving deepfakes with sexual content, portraying either famous persons or unknown individuals.[26] Regardless of the anonymity or not of the person depicted through the use of deepfakes, this constitutes a scenario in which undeniable damage to personal image occurs; precisely for this reason, in some jurisdictions, the regulatory treatment of such deepfakes is diverse.

A study conducted by Thorn, an organization aimed at preventing the sexual abuse of minors, showed that within a research group of 1,200 adolescents, 1 in 10 (10%) reported personally knowing someone whose image was used in a deepfake with sexual content. Moreover, 1 in 17 adolescents (6%) stated having been a direct victim of a deepfake with sexual content.[27] In other words, deepfakes are a tool whose

---

Luiz Inácio Lula da Silva being booed at an event in Uberlândia/MG (available at https://www.cnnbrasil.com.br/politica/carmen-lucia-manda-remover-videos-contra-lula-com-deep-fake/, Access: May 12, 2026).

23 Available at https://www.cbsnews.com/news/deepfakes-ai-fraud-elon-musk/, Access: May 12, 2026.

24 Available at https://uebermedien.de/83875/geschmacklose-geschichten-from-funke-media-germany/, Access: May 12, 2026.

25 Available at https://www.sport.de/news/ne5287557/formel-1--schumacher-familie-klagt-gegen-klatsch-magazin/, Access: May 12, 2026.

26 Available at https://www.bbc.com/portuguese/articles/c7299325zn3o and https://www.cbsnews.com/news/meta-facebook-sexualized-ai-deepfake-celebrity-images-spread/, Access: May 12, 2026.

27 Available at https://www.thorn.org/research/library/deepfake-nudes-and-young-people/, Access: May 12, 2026.

use can cause considerable harm if employed inadvertently.

Despite the extensive list of examples, deepfakes are not limited to such (illicit) cases, and the use of this technique for lawful purposes, especially in the media and entertainment industry, is becoming increasingly common.[28] In Brazil, an example of a deepfake that initially does not contain illegality is identified in a Volkswagen advertisement,[29] in which the late Brazilian singer Elis Regina was portrayed using artificial intelligence, singing alongside her daughter Maria Rita. Also noteworthy is the Mercado Livre advertisement[30] that recreated the voice of the father of the Brazilian footballer Zico as a tribute to the athlete.

Beyond these, there are other cases at national and international levels. In October 2024, it was announced that actor Peter Cushing, who died in 1994, would be "resurrected" to participate in a documentary about Hammer Film Productions, as the actor took part in 22 productions by the company. For this new project, the use of deepfakes was announced.[31]

Another relevant example was the "Dalí Lives" project, developed by The Dalí Museum in partnership with the advertising agency Goodby, Silverstein & Partners. In celebration of the 115th anniversary of Salvador Dalí's birth, the artist's image was recreated through deepfake technology to interact with museum visitors, who could ask questions about his life and receive authentic and new answers, not merely reproductions of previous images.[32]

The (non)existence of illicit purposes does not alter the subject matter of regulation. On the contrary, it demonstrates that a simple prohibition on the use of artificial intelligence for the creation of synthetic media would not address the issue efficiently. As Pawalec and Biess have argued,[33] complete bans on deepfakes "*are always*

---

28 MEDON, Filipe. O direito à imagem na era das *deepfakes*. *Revista Brasileira de Direito Civil – RBDCivil.* Belo Horizonte, v. 27, p. 251-277, jan./mar. 2021, p. 262.

29 Available at https://www.youtube.com/watch?v=aMl54-kqphE, Access: May 12, 2026.

30 Available at https://www.youtube.com/watch?v=DQEIKfl7VhI, Access: May 12, 2026.

31 Available at https://www.radiotimes.com/tv/documentaries/peter-cushing-ai-deepfake-hammer-documentary-newsupdate/, Access: May 12, 2026.

32 Available at https://thedali.org/press-room/dali-lives-museum-brings-artists-back-to-life-with-ai/, Access: May 12, 2026.

33 PAWALEC, Maria; BIEß, Cora. Deepfakes: Technikfolgen und Regulierungsfragen aus ethischer

associated with the risk of restricting freedom of expression in art, media, and the right to free expression of thought. Moreover, they may also impede the development of highly useful synthetic images and hinder research freedom."

As will be seen, the regulation of deepfakes must consider the entire landscape related to deepfakes, allowing their evolution and, simultaneously, their safe use.

3. POTENTIAL HARMS CAUSED BY DEEPFAKES AND GROUNDS FOR REGULATION

The dissemination of deepfakes, as can be inferred from the examples described above, most often results in the violation of the right of privacy of third parties (known in Brazil as the respect to one's image and honour).[34] In Brazilian legislation, the rights to image and honor are thus included among personality rights and are regulated by the provisions of Article 5, item X, of the Federal Constitution[35] and Article 20 of the Civil Code.[36] This topic is relevant to the present work insofar as deepfakes are inevitably associated with the image and honor of those who may be offended, making comprehensive legal protection of victims through regulation

und sozialwissenschaftlicher Perspektive. Baden-Baden: Nomos Verlagsgesellschaft, 2021, p. 6.

[34] KUGLER, Matthew; PACE, Carly. Deepfake Privacy: Attitudes and Regulation. Northwestern University Law Review, Vol. 116, 2021, p. 624.

[35] "Art. 5. All persons are equal before the law, without distinction of any kind; Brazilians and foreign nationals residing in the country are guaranteed the inviolability of the right to life, liberty, equality, security, and property, under the following terms: (...) X - the privacy, private life, honor, and image of persons are inviolable, and the right to compensation for material or moral damages resulting from their violation is guaranteed" (BRASIL. Constituição da República Federativa do Brasil de 1988. Brasília, DF. Available at https://www.planalto.gov.br/ccivil_03/constituicao/constituicao.htm, Access: May 12, 2026).

[36] "Art. 20. Except if authorized, or if necessary for the administration of justice or maintenance of public order, the disclosure of writings, the transmission of speech, or the publication, exhibition, or use of a person's image may be prohibited, upon their request and without prejudice to applicable compensation, if it affects their honor, good reputation, or respectability, or if intended for commercial purposes. Sole paragraph. In the case of a deceased or absent person, the legitimate parties to request such protection are the spouse, ascendants, or descendants." (BRAZIL. Law No. 10,406 of January 10, 2002. Institui o Código Civil. February 11, 2002, Brasília, DF. Available at https://www.planalto.gov.br/ccivil_03/leis/2002/l10406compilada.htm, Access: May 12, 2026).

necessary.

As Antônio Menezes Cordeiro teaches, these rights are included among personality rights: "*the fate given to the image is, in a way, a treatment given to the person themselves. The image thus appears on the stage of personality assets*."[37]

Regarding the right to one's image, Maria Helena Diniz[38] explains that it is the right to prevent the public exposure of someone's Portrait without their consent and the right not to have their personality materially or intellectually altered, causing damage to their reputation. Furthermore, according to Roxana Borges,[39] through the image right, the physical representation of a person is protected, whether fixed in photographs, films, videos, paintings, or other means that reproduce the person's face or parts of their body, physical signs, or gestures that may serve for identification and recognition.

Concerning the right to honor, also protected in Brazil under the mentioned Article 5, item X, of the Federal Constitution, Pontes de Miranda defines it as "*personal dignity, the feeling and awareness of being worthy, plus the esteem and moral consideration of others, which constitute the content of what is called honor*."[40]

Secondly, beyond harm to honor and image, in the current context, deepfakes are widely used in the electoral sphere. In this scenario, the regulation of deepfakes is justified from the perspective of protecting the democratic rule of law, preventing the use of technology to reinforce disinformation and electoral manipulation.[41]

On this subject, the "Panorama Político 2023" report, a study conducted by the Brazilian Federal Senate, indicated that at least 76% of the Brazilian population was

---

37 MENEZES CORDEIRO, Antonio. Tratado de direito civil. 4th ed., revised and updated, with collaboration by A. Barreto Menezes Cordeiro. Coimbra: Almedina, 2017, vol. 4, p. 255.

38 DINIZ, Maria Helena. Curso de Direito Civil Brasileiro, vol. 1, 21st ed., São Paulo: Saraiva, 2004, p. 127.

39 BORGES, Roxana Cardoso Brasileiro. Dos direitos da personalidade. In: LOTUFO, Renan; NANNI, Giovanni Ettore (Coords.). Teoria geral do direito civil. São Paulo: Atlas, 2008, p. 267.

40 MIRANDA, Pontes de. Tratado de Direito Privado. Vol. VII, Rio de Janeiro: Ed. Borsoi, 1971, p. 44.

41 TAVARES, Cláudio Mello. Artificial Intelligence and Deepfakes: Legal and Technological Challenges for the Integrity of the Democratic Process and Implications for the 2024 Municipal Elections. Revista Justiça Eleitoral em Debate, vol. 14, no. 1, 2025, p. 52.

exposed to false political information in the second half of 2023.[42] This fact demonstrates that the growth and popularization of generative artificial intelligence – which has become highly accessible – have introduced a new factor to be considered in this context, given the opportunities and risks this tool brings to democracy.

This is not a problem restricted to the Brazilian electoral process, as there are already several examples of the misuse of deepfakes in elections worldwide. Incidents include, among others, the United States, where a video circulated in which former U.S. President Barack Obama calls his opponent, Donald Trump, "a complete idiot", [43] and South Africa, where a deepfake video showed American rapper Eminem endorsing the Economic Freedom Fighters party and criticizing the African National Congress.[44] In Argentina, during the 2023 electoral campaign, then-candidates Javier Milei and Sergio Massa widely shared AI-generated images and videos in their campaigns, both for self-promotion and to attack the opposing candidate.[45]

Thus, the regulation of deepfakes, beyond the offense to the image and honor of the person portrayed – constitutionally guaranteed personality rights in Brazil –, is essential for the electoral process itself and the protection of the democratic rule of law.

Furthermore, deepfakes can be employed in contexts of fraud, causing financial harm to the respective victims, as exemplified by the video that circulated on social media featuring businessman Elon Musk promoting a crypto-asset investment, which ultimately turned out as a scheme for the misappropriation of third-party assets, as mentioned above.

A report produced by Sumsub Identity Verification[46] indicated that in 2023 Brazil had the highest number of fraud cases involving the use of deepfakes in Latin Ameri-

---

42 Available at https://www12.senado.leg.br/institucional/datasenado/publicacaodatasenado?id=panorama-politico-2023, Access: May 12, 2026.

43 Available at <https://www.businessinsider.com/obama-deepfake-video-insulting-trump-2018-4>, Access: May 12, 2026.

44 Available at https://africacheck.org/fact-checks/meta-programme-fact-checks/will-real-slim-shady-please-stand-eminem-video-endorsing, Access: May 12, 2026.

45 Available at https://www.nytimes.com/2023/11/15/world/americas/argentina-election-ai-milei-massa.html, Access: May 12, 2026.

46 Platform offering verification and authentication services for fraud prevention.

ca. According to the study, the majority of fraud cases in the country involving deepfakes that year were linked to crypto-assets, fintech companies, and the video game industry.

Finally, there is ongoing concern regarding this issue in the entertainment industry. Through deepfakes, actors can be easily replicated in cinematographic works, for example, without having actually participated in the filming. In 2023, there was a Hollywood actors' strike, one of the grounds of which was the use of deepfakes in the industry,[47] demonstrating a growing and significant concern.

As can be seen, there are various harms arising from deepfakes that justify their regulation. Based on different scenarios and priorities, China, the United States, and the European Union have adopted diverse measures to address the issue, which provide lessons for consolidating regulation of the matter in Brazil.

4. DEEPFAKES REGULATION

4.1. Deepfakes Regulation in China

China was one of the pioneers in creating specific laws to regulate deepfake technology and content generated by artificial intelligence. The country adopts a proactive approach to dealing with risks associated with misinformation, non-consensual pornography, and other illegal activities that can be facilitated by deepfakes.

The need for regulation in the country arose with the popularization of applications that allowed the creation of deepfakes in a simple and accessible manner. In 2019, an application called "ZAO" gained popularity in China by allowing users to swap faces in photos and videos,[48] and its popularity quickly sparked discussions about the app's data collection practices.

In this context, the Cyberspace Administration of China ("CAC"), responsible for

---

47 BLANCHET, Atahualpa. Hollywood's stand against AI: a blueprint for collective bargaining in the digital age. Available at: https://www.equaltimes.org/hollywood-s-stand-against-ai-a?lang=en. Access: May 15, 2026.

48 Available at: https://www.techtudo.com.br/noticias/2019/09/aplicativo-zao-usa-deepfake-para-criar-videos-e-viraliza-na-china.ghtml. Access: May 15, 2026.

national internet regulation, published several documents within the first three months of the app's launch, discussing the need to regulate the matter in the country.[49] As a consequence, an update to the so-called Administrative Provisions on the Management of Online Audio-Visual Information Service was published in November 2019.[50]

The legislation established new rules and responsibilities both for providers of audio and video information services and for users of these services. Regarding deepfakes, it was established, as a general rule, the illegality of publishing deepfakes without clear identification as such and the prohibition of using deep synthesis tools for producing fake news.[51]

With the continued growth and popularization of applications allowing the creation of deepfakes in the country, new specific rules on the topic were published in November 2022, the Provisions on the Administration of Deep Synthesis Internet Information Services (Deep Synthesis Provisions).[52] The object is the regulation of deep synthesis technology and responsibilities were established for providers of deep synthesis services and their technical supports concerning data security, personal information protection, transparency, and technical security.[53]

The legislation prohibits the production and dissemination of deepfakes with illegal content (Article 6), mandates obtaining consent from the person whose image will be distorted by means of deepfake, as well as confirming the real identity of the person using the technology (Article 9), and also obliges providers to improve technical management and regularly review, assess, and validate the mechanism and logic of their algorithms (Article 10).

The legislation, reinforcing the Administrative Provisions on Internet Audio-

---

49 GENG, Yinuo. Comparing Deepfake Regulatory Regimes in the United States, the European Union, and China. Georgetown Law Technology Review, 157, 2023, p. 167.

50 Available at https://www.chinalawtranslate.com/en/provisions-on-the-management-of-online-a-v-information-services/. Access: May 15, 2026.

51 GENG, Yinuo. Op. cit., p. 168.

52 Available at: https://www.chinalawtranslate.com/en/deep-synthesis/. Access: May 15, 2026.

53 ZHANG, Laney. China: Regulation of Artificial Intelligence. Research Paper. Washington: Law Library of Congress, 2023, p. 4.

Visual Program Service, further requires that AI-generated content, such as deepfakes, be clearly identified by adding visible watermarks or other labels informing the public that the content has been synthesized (Article 16).

It is evident that the legislation aims to regulate the manner of use of technologies that enable the creation of deepfakes. There is no focus on the outcomes of such use; rather, the focus on the underlying technology, the requirements for consent and real user verification, and the obligation for platforms to identify deepfakes demonstrate that the regulation primarily targets the initial phase of the deepfake lifecycle (creation), as well as the circulation of the material.

The regulation of deepfakes in the country aligns with the objective of maintaining social stability.[54] To achieve this regulatory goal, the Chinese legislator placed service providers that facilitate the creation of deepfakes at the center of regulation, requiring them to assume responsibility for the legality of deepfakes throughout their lifecycle.[55]

China assigns broad responsibilities to providers, especially in complying with more stringent requirements such as security assessments and the obligation to identify deepfakes. This restrictive regulatory environment imposes significant challenges, requiring providers to navigate a complex context of legal and technical demands to ensure compliance with established norms.

4.2. Deepfakes regulation in the European Union

The European Union adopts a multifaceted approach to regulating deepfakes, integrating them into a broader legal framework aimed at controlling the use of artificial intelligence and ensuring safety and transparency in the digital environment. The main legislative instruments addressing deepfakes are the Artificial Intelligence Regulation (AI Act) and the Digital Services Act (DSA), complemented by the

---

54 HINE, Emmie; FLORIDI, Luciano. New deepfake regulations in China are a tool for social stability, but at what cost? *Nature Machine Intelligence*, *4*(7), p. 608.

55 ZHENG, Guan; SHU, Jinchun; LI, Kaigeng. Regulating deepfakes between Lex Lata and Lex Ferenda – A comparative analysis of regulatory approaches in the U.S., the EU and China. *Crime, Law and Social Change, 83:1*, 2025, p. 13.

General Data Protection Regulation (GDPR), addressing different phases of the deepfake lifecycle.[56]

The AI Act is the EU's most direct legislation concerning deepfakes and other forms of deep synthesis. It adopts a risk-based approach to AI systems, dividing risks into four levels: minimal, low/limited (or with special transparency requirements), high, and unacceptable.[57]

The legislation does not explicitly use the term deepfake as a separate risk category but refers to AI systems that generate or manipulate audio, video, or image content. These systems are generally classified as limited risk, which entails transparency obligations.[58]

At the technology provision stage, providers and users of AI systems that generate or manipulate deepfakes must inform that the content was artificially generated or altered (Article 50). However, exceptions already provided by law include that labeling deepfake content is not required when done for law enforcement purposes, provided adequate safeguards are in place, or for artistic, satirical, or entertainment purposes, when the use does not mislead and information is appropriately provided.[59]

Furthermore, also at the initial phase of the deepfake lifecycle, the legislation introduces the concept of general-purpose AI models with systemic risk. Since AI systems such as ChatGPT – which may fall into this category – are capable of generating misleading texts or images, providers of these models, especially when used to

---

56 ZHENG, Guan; SHU, Jinchun; LI, Kaigeng. *Ob. Cit*., p. 8.

57 VOIGT, Paul; HULLEN, Nils. The EU AI Act: Answers to Frequently Asked Questions. New York: Springer Nature, 2024, p. 4.

58 EUROPEAN UNION. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act) (Text with EEA relevance), March 13, 2024. Available at: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32024R1689. Access: May 20, 2026.

59 MEDING, Kristof; SORGE, Christoph. What constitutes a Deep Fake? The blurry line between legitimate processing and manipulation under the EU AI Act. CSLAW '25: Proceedings of the 2025 Symposium on Computer Science and Law, pp. 152–159, 2025, p. 156.

create deepfakes, must fully comply with applicable legal requirements.[60]

Similarly, in certain contexts, the use of deepfakes may be considered high risk if integrated into AI systems classified as high risk under the AI Act (e.g., systems for managing critical infrastructures, credit assessment, or law enforcement). In such cases, obligations are much more stringent, including conformity assessments, risk management systems, and human oversight.

Likewise, deepfakes may fall within prohibited practices, such as the use of AI to exploit individuals' vulnerabilities, in which case their use is forbidden. As can be seen, generally there is a rule solely related to transparency duties, and since the regulation addresses AI as a whole, deepfakes may fall under various risk levels depending on their purpose.

Secondly, from the perspective of data protection, the GDPR is fundamental when deepfakes involve the use of personal data, especially biometric data (such as facial images or voice). Given the GDPR's broad definition of data processing, whether an AI system uses personal data for training or uses such data to generate specific deepfakes, both scenarios are considered personal data processing under the GDPR framework.

Article 6 of the GDPR[61] establishes six legal bases for processing personal data, requiring that at least one be present for the processing to be considered lawful.[62] In the specific case of deepfakes, the most relevant legal bases are usually the data subject's informed consent and the legitimate interests of the controller or third parties. When it comes to informed consent, if the deepfake replaces the image or identity of one data subject with another, consent must be obtained from both parties to ensure GDPR compliance.

---

60 BLOCK, Martina. Journal of European Consumer and Market Law; Vol. 13, Ed. 4, pp. 184–192, 2024, p. 186.

61 EUROPEAN UNION. Regulation (EU) 2016/679 of the European Parliament and of the Council. General Data Protection Regulation, April 27, 2016. Available at: https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679. Access: May 20, 2026.

62 GONZÁLEZ, Elena Gil; Hert, Paul de. Understanding the legal provisions that allow processing and profiling of personal data—an analysis of GDPR provisions and principles. ERA Forum 19, 597–621, 2019, p. 599.

Furthermore, the GDPR grants individuals rights over their data, including the right of access, rectification, erasure, and objection to processing. This means that if a person's image or voice is used in a deepfake without the existence of one of the legal basis, they may request its removal. It is noteworthy, however, that such provision may be ineffective given the frequent difficulty in identifying the author for the imposition of such obligations.[63]

Finally, at the content dissemination stage, the general rules of the DSA apply to deepfakes. The regulation aims "*to contribute to the proper functioning of the internal market for intermediary services by setting out harmonised rules for a safe, predictable and trusted online environment*"[64] and regulates the activity of hosting providers in cases of dissemination of illicit content published by third parties. Thus, whenever deepfakes containing illicit content are published on digital platforms (hosting providers), the notice and action system provided for in the aforementioned legislation [65] shall apply.

According to this model, providers are, at first, exempt from liability for illegal content published by third parties. This exemption applies as long as these platforms are not aware of the illegal activity or content based on objective facts or promptly take measures to remove or restrict access as soon as they become aware.

Providers must make available mechanisms for reporting illegal content and, once a report is received, must act to analyze it and promptly remove it, including deepfakes that fall within this scope. The broad European Union legislation, therefore, regulates deepfakes in a general maner throughout their lifecycle. At the moment of creation and application of the technology, the provisions of the AI Act apply, which initially impose only transparency obligations. Regarding the data access phase, the GDPR applies, and concerning dissemination, the DSA may be

---

63 ZHENG, Guan; SHU, Jinchun; LI, Kaigeng. Op. cit., p. 9.

64 EUROPEAN UNION. Regulation (EU) 2022/2065 of the European Parliament and of the Council on a Single Market for Digital Services and amending Directive 2000/31/EC, 19 Oct. 2022. Available at: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32022R2065. Access: May 20, 2026.

65 WOLTERS, Pieter; GELLERT, Raphael. Towards a better notice and action mechanism in the DSA. 14 Journal of Intellectual Property, Information Technology and Electronic Commerce Law 403, 2023, p. 404.

applied.

The application of generic regulation to the particular case, on the one hand, allows for stable legislation that keeps pace with technological evolution and, on the other hand, leaves gaps regarding the specific regulation of deepfakes. Among the criticisms of the European Union regulations are the absence of an objective definition of what constitutes a deepfake, the overlap between obligations of providers and users, and the lack of clear exceptions for certain cases.[66]

4.3. Deepfakes regulation in the United States

The regulation of deepfakes in the United States has developed on two main levels: federal and state. Recently, there has been significant progress at the federal level, but most specific legislation has come from the states.

Since 2018, members of the U.S. House of Representatives and Senate have introduced various legislative proposals aimed at regulating deepfakes. These proposals sought to penalize creators and platforms that distribute deepfakes for illicit activities, impose transparency obligations for deepfake content, assess risks associated with the technology, and protect individuals from deepfakes exploiting intimate relationships.[67] Despite these efforts, none of these bills were enacted.[68]

Only in 2025 was the so-called Take It Down Act enacted.[69] This federal law criminalizes the distribution of non-consensual intimate images, including AI-generated deepfakes. The legislation imposes penalties of up to three years' imprisonment and fines on those responsible, in addition to requiring digital platforms to

---

66 MEDING, Kristof; SORGE, Christoph. Op. Cit., p. 156.

67 The regulation of deepfakes in the United States has developed on two main levels: federal and state. Recently, there has been significant progress at the federal level, but most specific legislation has come from the states.

68 GUZMAN, Jonathan. Saving face: Lessons from the DMCA for combating deepfake pornography. Gonzaga Law Review, 58(1), 109–138, 2022, p. 111.

69 UNITED STATES OF AMERICA. Tools to Address Known Exploitation by Immobilizing Technological Deepfakes on Websites and Networks Act (Take It Down Act). Senate Bill 146, 119th Congress, 1st sess., Public Law No. 119-12, § 1 et seq., enacted May 19, 2025. Available at: https://www.congress.gov/bill/119th-congress/senate-bill/146/text. Access: May 20, 2026.

remove such publications within 48 hours after notification by the victim and to take measures to eliminate other known copies.

The legislation regulates deepfakes in a sectoral manner and constitutes an important measure to combat revenge pornography using deepfakes. Nevertheless, the divergent approach to the subject in the United States is clear, since there is no intention to regulate deepfakes as a whole or the underlying technology, but only the acts of individuals who disseminate illicit content and their respective consequences, specifically regarding revenge pornography.

In the country, the approval of generic legislation on deepfakes would likely be impeded by the First Amendment to the U.S. Constitution, which addresses, among other things, freedom of expression.[70] For some jurists, who support their position based on Supreme Court rulings, deepfakes constitute a form of expression protected by the First Amendment.[71] Precisely for this reason, regulation of the matter in the United States, both at the federal and state levels, is limited to specific situations such as pornography and electoral contexts.[72]

At the state level, some states already have laws to curb deepfakes in elections. Noteworthy is California legislation: Assembly Bill 2839 prohibits the dissemination of deepfakes within 120 days before elections.[73] In the same state, Assembly Bill 602

---

70 "Congress shall make no law respecting an establishment of religion, or prohibiting the free exercise thereof; or abridging the freedom of speech, or of the press; or the right of the people peaceably to assemble, and to petition the Government for a redress of grievances." UNITED STATES OF AMERICA. Constitution of the United States of America, First Amendment. Ratified 15 Dec. 1791. Available at: https://constitution.congress.gov/browse/amendment-1/#:~:text=First%20Amendment%20Fundamental%20Freedoms,for%20a%20redress%20of%20grievances. Access: May 20, 2026.

71 FEENEY, Matthew. Deepfake Laws Risk Creating More Problems Than They Solve, Regulatory Transparency Project of The Federalist Society, 2021. Available at: https://regproject.org/paper/deepfake-laws-risk-creating-more-problems-than-they-solve/. Access: May 20, 2026.; BLITIZ, Jonathan. Deepfakes and Other Non-Testimonial Falsehoods: When is Belief Manipulation (Not) First Amendment Speech. 23 Yale Journal of Law and Technology, 2020, p. 160.

72 GENG, Yinuo. Op. Cit., p. 165.

73 CALIFORNIA (State). Assembly Bill No. 2839, An act to amend Section 35 of the Code of Civil Procedure, and to add Section 20012 to the Elections Code, relating to elections, and declaring the urgency thereof, to take effect immediately. Available at: https://leginfo.legisla-

creates a private right of action against those who intentionally distribute deepfakes of intimate or sexual photographs and videos.[74] Other states follow the same trend, with legislation concerning deepfakes falling within the realms of pornography[75] or electoral matters, with labeling requirements for deepfake publications being common in the latter.[76]

Although there is growing recognition of the need for a unified approach to deepfakes, especially in scenarios of disinformation and national security, legislative fragmentation among states remains a characteristic of U.S. regulation on the subject. The enactment of the federal Take It Down Act represents an important step toward harmonizing legislation regarding non-consensual sexual deepfakes, but regulation of other malicious uses of deepfakes remains a challenge and an area of ongoing development, especially in light of the First Amendment.

4.4. Regulatory Models in Summary

From the analysis of deepfake regulation in China, the European Union, and the United States, some similarities and differences can be observed. From the perspective of the deepfake lifecycle, while China and the European Union focus on the initial stages (technology, creation of deepfakes, and dissemination), with legislation aimed at platforms and individual creators of deepfakes, the United States proposes later measures focused on the public and the victim.

When analyzing the regulations from a sectoral perspective, it is noticeable that the United States, by promoting specific legislation in the electoral field and concerning pornography, creates sectoral regulation of the matter. Europe and China do not present sectoral regulation but promote regulation based on risk analysis.

---

ture.ca.gov/faces/billNavClient.xhtml?bill_id=202320240AB2839. Access: May 20, 2026.

74 CALIFORNIA (State). act to amend Sections 65940.1 and 66019 of, and to add Section 66016.5 to, the Government Code, and to add Section 50466.5 to the Health and Safety Code, relating to land use, approved 3 Oct. 2019. Available at: https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=202120220AB602. Access: May 20, 2026.

75 Virginia, California, Texas, Indiana, Hawaii, Georgia, Florida, Louisiana, Minnesota, New York, North Carolina, South Dakota, Utah, Washington.

76 Texas, California, Minnesota, Washington, Michigan, New York, Indiana, Utah, Wisconsin.

While in European Union legislation risk-based regulation is evidente – given the division into risk categories – China's position can be perceived from the fact that the legislation does not bring specific regulation of the technology known as deepfake, but rather of deep synthesis applications, which demonstrates an attempt at more generic regulation focused on the consequences of misuse of the technology, through a risk weighting system.

Finally, a third way to organize this analysis is through differentiation between general or more specific legislation. In this regard, the United States regulate the matter specifically, considering deepfakes as a unique technology. Similarly, in China, the promulgation of the Deep Synthesis Provisions demonstrates that the country adopts a form of regulation that considers the technology individually. However, as stated earlier, it cannot be overlooked that such legislation is inserted within a broader and more general context of artificial intelligence regulation, forming a hybrid model in the country.

On the other hand, in the European Union, there is one (or more) general regulation applicable to deepfakes, but which addresses artificial intelligence as a whole. A positive aspect of this form of regulation is the possibility of learning from various scenarios, enabling the addressing of specific issues more efficiently and possibly more durably.[77]

5. PATHS FOR REGULATION OF THE MATTER IN BRAZIL

In Brazil, despite growing concern regarding the subject, there is still no specific or even general legislation addressing deepfakes. Nevertheless, legislative proposals aimed at regulating this type of technology have been discussed before the National Congress.

One important proposal is the Projeto de Lei nº 2.338,[78] which proposes the

---

77 GENG, Yinuo. Op. Cit., p. 174.

78 BRAZIL. Federal Senate. *Projeto de Lei nº 2.338, de 2023. Dispõe sobre o uso da inteligência artificial no Brasil, com base na dignidade humana, na proteção dos direitos fundamentais e no desenvolvimento tecnológico.* Brasília, DF: Federal Senate, 2023. Available at: https://www25.senado.leg.br/web/atividade/materias/-/materia/157233. Access: May 24, 2026.

regulation of artificial intelligence in Brazil. Although it does not expressly mention deepfakes, the proposal intends to regulate artificial intelligence generically, in a risk-based model, so that, regardless of explicit indication, deepfakes would be encompassed by the regulation. Specifically addressing deepfakes that use the image of a person after their death, the proposal for revision and updating of the Civil Code[79] provides for a provision allowing the creation of images of deceased persons through artificial intelligence, provided that certain conditions are met, including the prior and express informed consent of the legal heirs or representatives of the deceased.

This article does not intend to map all bills presented involving the subject, especially since there are several others, but to highlight that there is legislative attention to the matter, making clear that, until the present moment, there is no existing regulation concerning artificial intelligence or specifically deepfakes. Nevertheless, there are some sectoral norms that address the topic.

In the electoral context, in March 2024, the Superior Electoral Court (TSE) published Resolution No. 23,732, which provides rules on electoral propaganda and introduces norms for the use of artificial intelligence in electoral campaigns and political communication on social media platforms, to be observed by candidates, parties, and application providers. Among these, a total prohibition on the dissemination of deepfakes in campaigns was established.[80]

Furthermore, on April 25, 2025, a law was enacted that increases the penalty for

---

[79] BRAZIL. Federal Senate. *Projeto de Lei nº 4, 2025. Amends Law No. 10,406, of January 10, 2002 (Civil Code)*, and related laws. Brasília, DF: Federal Senate, 2025. Available at: https://www25.senado.leg.br/web/atividade/materias/-/materia/166998. Access: May 24, 2026.

[80] Art. 9º-C The use, in any form or modality, of fabricated or manipulated content to disseminate notoriously false or decontextualized facts with the potential to cause harm to the balance of the election or the integrity of the electoral process is prohibited in electoral propaganda. § 1 The use of synthetic content in audio, video, or a combination thereof, generated or digitally manipulated, even with authorization, to create, replace, or alter the image or voice of a living, deceased, or fictitious person (deepfake), to harm or favor a candidacy, is prohibited. BRAZIL. Superior Electoral Court (TSE). Resolução nº 23.732, February 27, 2024. Provides rules on electoral propaganda and introduces norms for the use of artificial intelligence in electoral campaigns and political communication on social media platforms. Brasília, DF: Superior Electoral Court, 2024. Available at: https://www.tse.jus.br/legislacao/compilada/res/2024/resolucao-no-23-732-de-27-de-fevereiro-de-2024. Access: May 24, 2026.

crimes of violence against women involving the use of artificial intelligence, a spectrum under which deepfakes are categorized.[81]

As observed, these are highly specific regulations, leaving room for discussion and the creation of legislation that satisfactorily governs the matter in the country, using international experiences as examples. Such a study is important given that the country is particularly affected by the growth in the use of deepfakes.[82]

In this context, one of the advantages of new technologies is highlighted: their rapid and *pari passu* dissemination across various jurisdictions allows countries to learn from one another regarding the optimal regulation of matters involving new technologies.

Thus, in the present article, based on the international regulations previously analyzed, it is proposed that the ideal regulation of the matter in Brazil involves the creation of specific legislation addressing deepfakes as a unique technology. With the existence of explicit regulation, greater effectiveness in addressing the issue is expected, especially by imposing special obligations such as the duty to identify content containing deepfakes, as is done in China, in addition to providing for explicit and justified exceptions, as practiced in the European Union, and clear coercive measures, as foreseen in the American Take It Down Act.

Despite the specific regulation of deepfakes, the necessity that such legislation be part of a broader strategy aimed at regulating artificial intelligence as a whole is not dismissed. The objective is to enable the growth of the technology in a general context, which often implies the need for special attention to certain nuances.

Moreover, it is proposed that, although sectoral regulation of existing matters –

---

[81] BRAZIL. Lei nº 15.123, April 24, 2025. Amends art. 147-B of Decree-Law No. 2,848, December 7, 1940 (Penal Code), to establish an aggravating circumstance for the crime of psychological violence against women when committed using artificial intelligence or any other technological resource that alters the victim's image or sound. Brasília, DF: Presidency of the Republic, 2025. Available at: https://www.planalto.gov.br/ccivil_03/_ato2023-2026/2025/lei/L15123.htm. Access: May 24, 2026.

[82] Data from the Identity Fraud Report 2023 by Sumsub demonstrate that between 2022 and 2023 Brazil registered an 830% increase in circulating deepfakes across all sectors. Available at: https://static.poder360.com.br/2025/02/Identity-fraud-report-sumsub-2023.pdf. Access: May 24, 2026.

such as electoral law and violence against women – is important, the creation of comprehensive regulation concerning the technology itself is necessary, not limited to specific use cases.

Additionally, focus on the initial and intermediate stages of the deepfakes' life cycle is important to promote equitable treatment at these stages and to prevent the uncontrolled increase in the use of such technology. In this regard, it is essential that the obligations of the involved parties and the corresponding penalties be clearly defined.

The international experience analyzed in this article is therefore of great value for establishing an appropriate regulatory regime in Brazil. It is expected that, based on the prior experience of other countries such as China, the European Union, and the United States, the Brazilian legislator may find the best way to regulate the matter domestically, having presented some normative characteristics that, in the author's view, are relevant for such regulation.

6. CONCLUSION

Deepfakes, defined as realistic manipulations of media through artificial intelligence, can generate fictitious and highly convincing content. Their capacity to produce new media from pre-existing models that are hardly identifiable to the naked eye renders their regulation imperative, encompassing the technology itself, as well as the circulation and access to the content.

The potential harms are vast, including violations of image and honor rights, infringement of electoral process integrity by promoting misinformation, and violations of property rights in cases of fraud. It is precisely the existence of such potential harms that justifies increased attention to the regulation of deepfakes.

China, a pioneer in specific regulation of the matter, focuses on regulating the technology and holding providers accountable, requiring consent, identity verification, and clear labeling of content generated by artificial intelligence. Although this may impose a greater burden on providers, the country presents a cohesive regulatory framework capable of addressing the life cycle of deepfakes at the initial and intermediate stages.

The European Union, in turn, adopts a multifaceted approach with the AI Act (regulating AI systems based on risk and requiring transparency), the GDPR (protecting personal data in deepfakes), and the Digital Services Act (holding hosting providers liable for illicit content). Regulation of the matter in the European continent is conducted in a generic manner, encompassing deepfakes within the broader spectrum of artificial intelligence.

In the United States, regulation is divided between federal and state levels, with the Take It Down Act criminalizing the distribution of non-consensual intimate images and states such as California enacting specific laws against deepfakes in electoral campaigns and revenge pornography. In the country, freedom of expression (First Amendment) constitutes a significant limitation on the scope of general legislation on the subject.

In Brazil, despite growing legislative attention, there is still no specific legislation for deepfakes, which allows for a more detailed study of the experiences of other jurisdictions such as China, the European Union, and the United States. Learning from international experiences is crucial for Brazil to develop effective and satisfactory regulation.

In the concrete case, in light of the jurisdictions listed above, an ideal regulation is understood to be one that provides specific provisions but is embedded within a general context of artificial intelligence regulation, which furnishes tools for identifying deepfakes and, at the same time, reinforces the need for compliance with legislation through coercive measures. Furthermore, it is fundamental that the legislation be comprehensive regarding the technology itself, not limited solely to specific uses, and that it focus especially on the initial and intermediate stages of the deepfakes' life cycle, allowing technological evolution in a balanced manner.

REFERENCES

ARBEL, Yonathan; MATTHEW, Tokson; LIN, Albert. Systemic Regulation of Artificial Intelligence. *Arizona State Law Journal*. Phoenix: Arizona State University. v. 56, n. 2, p. 545-620, Summer, 2024.

BLANCHET, Atahualpa. *Hollywood's stand against AI: a blueprint for collective bargaining in the digital age*. Available at: https://www.equaltimes.org/hollywood-s-stand-against-ai-a?lang=en.

BLITZ, Jonathan. Deepfakes and Other Non-Testimonial Falsehoods: When is Belief Manipulation (Not) First Amendment Speech. *23 Yale Journal of Law and Technology*, 2020.

BLOCK, Martina. *Journal of European Consumer and Market Law*; Vol. 13, Ed. 4, p. 184-192, 2024.

BORGES, Roxana Cardoso Brasileiro. Dos direitos da personalidade. In: LOTUFO, Renan; NANNI, Giovanni Ettore (Coord.). *Teoria geral do direito civil*. São Paulo: Atlas, 2008. BOSTROM, Nick. *Superintelligence: Paths, Dangers, Strategies.* Translated by Aurélio Antônio Monteiro, Clemente Gentil Penna, Fabiana Geremias Monteiro, Patrícia Ramos Geremias. Rio de Janeiro: DarkSide, 2018, p. 38.

BRASIL. *Constituição da República Federativa do Brasil de 1988*. Brasília, DF. Available at https://www.planalto.gov.br/ccivil_03/constituicao/constituicao.htm.

BRAZIL. *Lei nº 10.406 de 10 de janeiro de 2002*. Institui o Código Civil. February 11, 2002, Brasília, DF. Available at https://www.planalto.gov.br/ccivil_03/leis/2002/l10406compilada.htm.

BRAZIL. Federal Senate. *Projeto de Lei nº 2.338, de 2023. Dispõe sobre o uso da inteligência artificial no Brasil, com base na dignidade humana, na proteção dos direitos fundamentais e no desenvolvimento tecnológico.* Brasília, DF: Federal Senate, 2023. Available at https://www25.senado.leg.br/web/atividade/materias/-/materia/157233.

BRAZIL. Federal Senate. *Projeto de Lei nº 4, 2025. Amends Law No. 10,406, of January 10, 2002 (Civil Code)*, and related laws. Brasília, DF: Federal Senate, 2025. Available at https://www25.senado.leg.br/web/atividade/materias/-/materia/166998.

BRAZIL. Superior Electoral Court (TSE). *Resolução nº 23.732, February 27, 2024*. Provides rules on electoral propaganda and introduces norms for the use of artificial intelligence in electoral campaigns and political communication on social media platforms. Brasília, DF: Superior Electoral Court, 2024. Available at https://www.tse.jus.br/legislacao/compilada/res/2024/resolucao-no-23-732-de-27-de-fevereiro-de-2024.

BRAZIL. *Lei nº 15.123, April 24, 2025*. Amends art. 147-B of Decree-Law No. 2,848, December 7, 1940 (Penal Code), to establish an aggravating circumstance for the crime of psychological violence against women when committed using artificial intelligence or any other technological resource that alters the victim's image or sound. Brasília, DF: Presidency of the Republic, 2025. Available at: https://www.planalto.gov.br/ccivil_03/_ato2023-2026/2025/lei/L15123.htm.

CALIFORNIA (State). *Assembly Bill No. 2839*. An act to amend Section 35 of the Code of Civil Procedure, and to add Section 20012 to the Elections Code, relating to elections, and declaring the urgency thereof, to take effect immediately. Available at: https://leginfo.legislature.ca.gov/faces/billNavClient.xhtml?bill_id=202320240AB2839.

CALIFORNIA (State). *Assembly Bill No. 602*. An act to amend Sections 65940.1 and 66019 of, and to add Section 66016.5 to, the Government Code, and to add Section 50466.5 to the Health and Safety Code, relating to land use, approved 3 Oct. 2019. Available at: https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=201920200AB602.

CAJUEIRO, Daniel O.; CELESTINO, Victor R. R. *A Comprehensive Review of Artificial Intelligence Regulation: Weighing Ethical Principles and Innovation*. Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4950727.

CAZZANIGA, Martina. Una nuova tecnica (anche) per veicolare desinformazione: le risposte europee ai deepfakes. In: *Rivista di Diritto dei Media*, vol. 1, 2023.

CHESNEY, Bobby; CITRON, Daniele. *Deepfakes: a looming crisis for national security, democracy and privacy*. 107 California Law Review 1753, 2019.

DINIZ, Maria Helena. *Curso de Direito Civil Brasileiro*, v.1, 21st Ed., São Paulo: Saraiva, 2004.

EUROPEAN UNION. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act) (Text with EEA relevance), March 13, 2024. Available at: https://eur-lex.europa.eu/legal-content/PT/TXT/?uri=CELEX:32024R1689.

EUROPEAN UNION. Regulation (EU) 2016/679 of the European Parliament and of the Council. General Data Protection Regulation, April 27, 2016. Available at: https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679.

EUROPEAN UNION. Regulation (EU) 2022/2065 of the European Parliament and of the Council on a Single Market for Digital Services and amending Directive 2000/31/EC, October 19, 2022. Available at: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32022R2065.

FEENEY, Matthew. *Deepfake Laws Risk Creating More Problems Than They Solve, Regulatory Transparency Project of The Federalist Society*, 2021. Available at https://regproject.org/paper/deepfake-laws-risk-creating-more-problems-than-they-solve/.

FINNOCCHIARO, Giusella. The Regulation of Artificial Intelligence. *AI & Society*. New York: Springer Nature, v. 39, p. 1961-1968, April, 2023.

GENG, Yinuo. Comparing Deepfake Regulatory Regimes in the United States, the European Union, and China. *Georgetown Law Technology Review*, 157, 2023.

GONZÁLEZ, Elena Gil; HERT, Paul de. *Understanding the legal provisions that allow processing and profiling of personal data—an analysis of GDPR provisions and principles*. ERA Forum 19, 597–621, 2019.

GUZMAN, Jonathan. Saving face: Lessons from the DMCA for combating deepfake pornography. *Gonzaga Law Review*, 58(1), 109–138, 2022.

HINE, Emmie; FLORIDI, Luciano. New deepfake regulations in China are a tool for social stability, but at what cost? *Nature Machine Intelligence*, 4(7), 2022.

KUGLER, Matthew; PACE, Carly. Deepfake Privacy: Attitudes and Regulation. *Northwestern*

University Law Review, Vol. 116, 2021.

MEDING, Kristof; SORGE, Christoph. What constitutes a Deep Fake? The blurry line between legitimate processing and manipulation under the EU AI Act. CSLAW '25: Proceedings of the 2025. *Symposium on Computer Science and Law*, pp. 152–159, 2025.

MEDON, Filipe. O direito à imagem na era das *deepfakes*. *Revista Brasileira de Direito Civil – RBD Civil*. Belo Horizonte, vol. 27, pp. 251–277, Jan./Mar. 2021.

MENEZES CORDEIRO, António. *Tratado de direito civil*. 4th ed., revised and updated, with the collaboration of A. Barreto Menezes Cordeiro. Coimbra: Almedina, 2017.

MIRANDA, Pontes de. *Tratado de Direito Privado*. Vol. VII, Rio de Janeiro: Ed. Borsoi, 1971.

MITRA, Alakananda; MOHANTY, Saraju; KOUGIANOS, Elias. *The World of Generative AI: Deepfakes and Language Models*. Available at: https://doi.org/10.48550/arXiv.2402.04373, accessed June 10, 2025.

NARAYANAN, Arvind; KAPOOR, Sayash. *AI Snake Oil: What Artificial Intelligence Can Do, What It Can't, and How to Tell the Difference*. Princeton: Princeton University Press, 2024.

PAWALEC, Maria; BIEß, Cora. *Deepfakes: Technikfolgen und Regulierunsfragen aus ethischer und sozialwissenschaflitcher Perspektive*. Baden-Baden: Nomos Verlagsgesellschaft, 2021. RAMATIK, Amela. *Küsntliche Inteligenz im Rahmen der EU*. Johannes Kepler University Linz, 2023, p. 23. Available at: https://epub.jku.at/obvulihs/content/titleinfo/8631729/full.pdf;

SCHREIBER, Anderson; RIBAS, Felipe; MANSUR, Rafael. Deepfakes: regulação e responsabilidade civil. In: TEPEDINO, Gustavo; SILVA, Rodrigo da Guia (Coords.). *O Direito Civil na Era da Inteligência Artificial*. São Paulo: Thomson Reuters Brasil, 2020.

SEARLE, John. *Mind, Brain, and Science*. Translated by Artur Morão. Lisbon: Library of Contemporary Philosophy, 2000.

SHOAIB, Mohamed; WANG, Zefan; AHVANOOEY, Milada; ZHAO, Jun. Deepfakes, Misinformation, and Disinformation in the Era of Frontier AI, Generative AI, and Large AI Models. *International Conference on Computer and Applications* (ICCA), 2023.

SOLUM, Lawrence B. Legal Personhood for Artificial Intelligences. *North Carolina Law Review*. Chapel Hill: The University of North Carolina School of Law, vol. 70, no. 04, Apr. 1992.

STRAUB, Sebastian. *The New Legal Framework for Artificial Intelligence – An Overview of the Planned EU Regulation on AI*. Available at: https://www.wolterskluwer.com/de-de/expert-insights/der-neue-rechtsrahmen-fuer-kuenstliche-intelligenz, accessed June 15, 2025.

TAVARES, Cláudio Mello. Inteligência Artificial e Deepfakes: Desafios Jurídicos e Tecnológicos para a Integridade do Processo Democrático e as Implicações para as Eleições Municipais de 2024. In: *Revista Justiça Eleitoral em Debate*, vol. 14, no. 1, 2025.

UNITED STATES OF AMERICA. Tools to Address Known Exploitation by Immobilizing Technological Deepfakes on Websites and Networks Act (Take It Down Act). *Senate Bill 146*, 119th

Congress, 1st sess., Public Law No. 119-12, § 1 et seq., enacted May 19, 2025. Available at: https://www.congress.gov/bill/119th-congress/senate-bill/146/text.

UNITED STATES OF AMERICA. *Constitution of the United States of America, First Amendment*. Ratified December 15, 1791. Available at: https://constitution.congress.gov/browse/amendment-1/#:~:text=First%20Amendment%20Fundamental%20Freedoms,for%20a%20redress%20of%20grievances.

VOIGT, Paul; HULLEN, Nils. *The EU AI Act: Answers to Frequently Asked Questions*. New York: Springer Nature, 2024.

WOLTERS, Pieter; GELLERT, Raphael. Towards a Better Notice and Action Mechanism in the DSA. *Journal of Intellectual Property, Information Technology and Electronic Commerce Law*, vol. 14, p. 403, 2023.

ZHANG, Laney. *China: Regulation of Artificial Intelligence*. Research Paper. Washington: Law Library of Congress, 2023.

ZHENG, Guan; SHU, Jinchun; LI, Kaigeng. Regulating Deepfakes between Lex Lata and Lex Ferenda – A Comparative Analysis of Regulatory Approaches in the U.S., the EU and China. *Crime, Law and Social Change*, vol. 83, no. 1, 2025.

CODEX: PROSPECTS AND CHALLENGES IN THE AUTOMATED ANALYSIS OF CASE LAW

Guilherme Ramos de Morais
PhD candidate (UFMG)
Master of Laws (Unb)
Postgraduate Specialisation in Civil Procedure (UERJ)
Civil Lawyer

DOI: https://doi.org/10.59224/dti8.ch19

Abstract: The paper analyzes the development, normative and algorithmic structures, operational advantages, and challenges related to the Codex Platform, a Brazilian data lake of procedural information. The paper is divided as follows: introduction; platform history; normative basis; algorithmic structure; operational advantages; challenges related to its implementation; issues regarding information security and personal data protection. At the end, advantages and challenges related to the Codex are compiled. The report is based on a bibliographic review of specialized literature.

Keywords: Codex; Natural language processing; Information retrieval.

Table of contents: 1. Introduction; 2. Planning, development, and implementation of Codex; 3. Technical Features of Codex; 4. Discussion on the potential impacts of Codex on improving information retrieval; 4.1. Unified National Indexing; 4.2. Nationwide Semantic Retrieval; 5. Regulatory Challenges Related to the Implementation of Codex and Its Integration with AI Models; 5.1. Data Access and APIs; 5.2. Information Security; 5.3. Personal Data Protection; 6. Final remarks.

1. INTRODUCTION

The short story *The Library of Babel*, written by the Argentine Jorge Luis Borges in 1941, presents an unusual library. A building composed of an indefinite, indeed, infinite, number of hexagonal galleries, from which one can observe infinite floors extending both upward and downward. Each gallery has twenty shelves with five bookcases; each bookcase contains thirty-two uniform books; each book has four hundred and ten pages; each page has forty lines; each line contains eighty black letters. As it is infinite, everything that has been or will be written is necessarily contained in the library. Beyond philosophical analyses, the library has a unique characteristic: it is a unified repository of infinite knowledge.

This type of document-based knowledge management has both advantages and disadvantages. Its greatest advantage is the possibility of gathering all available knowledge in a single place. However, this very advantage also introduces an inherent challenge: the volume of information makes it difficult to retrieve what is, indeed, relevant. Moreover, this centralization also turns the documentary repository into a single target for malicious attacks.

Similar to the Library of Babel, in 2020 the Brazilian National Council of Justice (*Conselho Nacional de Justiça)* (CNJ) began implementing the Codex Platform. It is a data lake of procedural information, that is, a unified repository for the document management of electronic court cases. In april 2026, more than 97% of all Brazilian procedural data had already been uploaded to the platform.[1] Due to its robustness, Codex has the potential to generate positive applications for the structuring and retrieval of information. However, the limited public disclosure about it raises ques-

[1] BRAZIL. National Justice Council. *Painel de acompanhamento da Plataforma Codex*. Brasília, [n.d.-a]. Available at: https://paineisanalytics.cnj.jus.br/single/?appid=de543e38-1c79-412e-b7b2-51624b55349d&sheet= 2c946c67-9efe-4e23-a74c-d1c4792fb0bb. Accessed on: Oct. 24, 2025.

tions regarding its purpose, as well as concerns about data accessibility and information security.

Having set out these preliminary considerations, this chapter analyzes the development, normative and algorithmic structure, operational advantages, and challenges related to the Codex Platform implementation. It also examines regulatory aspects concerning the application of natural language processing models in this type of database. The structure of this chapter follows this same linear logic: introduction; background of the Codex Platform; analysis of its normative framework; analysis of its algorithmic structure; discussion of operational advantages associated with its implementation, particularly the possibility of creating a unified national indexer and a semantic retrieval database designed for the legal domain; challenges related to its implementation, especially regulatory issues concerning the processing of the data contained in this database, including questions regarding data access; information security (in infrastructural terms); and personal data protection. Finally, the chapter presents its concluding remarks, compiling the main prospects and challenges regarding the Codex Platform.

The methodology consists of a literature review on the subject, as well as an analysis of the normative acts related to the platform. To ensure ethical transparency in the use of artificial intelligence, it should be noted that grammatical revision and suggestions for improving the original wording were carried out using generative artificial intelligence tools, such as GPT-5.3 and Gemini AI Pro 3.0. No generative artificial intelligence was used as a substitute for the author or for the original development of ideas.

2. PLANNING, DEVELOPMENT, AND IMPLEMENTATION OF CODEX

According to the National Council of Justice, Codex is the result of a partnership between the Council and the Court of Justice of The State of Rondônia (*Tribunal de Justiça de Rondônia, TJRO*), established around 2020.[2] However, collaboration be-

[2] BRAZIL. National Justice Council. *Plataforma Codex*. Brasília, [n.d.-b]. Available at:

tween the Counsil and TJRO on technological solutions for judicial data management dates back to 2018, when a technical cooperation agreement was signed allowing the Council to use TJRO's Sinapses System nationwide.[3] Despite these facts, the CNJ officially states, within the portfolio of the Justice 4.0 Program (*Programa Justiça 4.0*), that Codex was created in 2022, in the following terms: *"The Codex platform is also part of the AI ecosystem. It was created in 2022 to be the official framework for collecting and storing all data from electronic lawsuits in Brazil."*[4]

This information contradicts Annex II of CNJ Ordinance No. 118, which, since April 13, 2021, specifically regulates the portfolio of information and communication technology solutions and digital services of the CNJ.[5] In that ordinance, Codex was already listed among the CNJ's technologies. Publicly available information about its origins effectively ends here: there is little reliable public information regarding its actual origin, the individuals and public bodies involved, or the financial resources underlying its development.

From a normative standpoint, not intending to exhaust the analysis of constitutional and infra-constitutional norms on the subject, it is useful to emphasize the context in which Codex emerged. In 2020, CNJ established the National Judicial Data Repository (DATAJUD) under CNJ Resolution No. 331/2020[6], whose data were still submitted by courts in a non-automated manner. The primary object of analysis

https://www.cnj.jus.br/sistemas/plataforma-codex/. Accessed on: Oct. 24, 2025.

3 BRAZIL. National Justice Council. *Inteligência artificial: parceria com tribunal de Rondônia aproxima o futuro*. Brasília, 2018. Available at: https://www.cnj.jus.br/inteligencia-artificial-parceria-com-tribunal-de-rondonia-aproxima-o-futuro/. Accessed on: Oct. 24, 2025.

4 Brasil. Conselho Nacional de Justiça. (2022a). Justice 4.0 Program. https://www.cnj.jus.br/wp-content/uploads/2022/05/justice-4-0-program.pdf

5 BRAZIL. National Justice Council. *Portaria nº 118, de 13 de abril de 2021*. Dispõe sobre o portfólio de soluções de tecnologia da informação e comunicação e serviços digitais do Conselho Nacional de Justiça. Brasília, 2021a. Available at: https://atos.cnj.jus.br/atos/detalhar/3866. Accessed on: Oct. 24, 2025.

6 BRAZIL. National Justice Council. *Resolução nº 331, de 20 de agosto de 2020*. Institui a Base Nacional de Dados do Poder Judiciário - DataJud como fonte primária de dados do Sistema de Estatística do Poder Judiciário - SIESPJ para os tribunais indicados nos incisos II a VII do art. 92 da Constituição Federal. Brasília, 2020e. Available at: https://atos.cnj.jus.br/atos/detalhar/3428. Accessed on: Oct. 24, 2025.

within DATAJUD consisted of procedural metadata (unified case numbering, pursuant to CNJ Resolution No. 65/2008[7]; codes from the Unified Procedural Table (TPU), pursuant to CNJ Resolution No. 46/2007[8]; parties' data, among others). Shortly thereafter, in CNJ Resolution No. 335/2020[9], the CNJ established the Brazilian Judiciary Digital Platform (PDPJ-Br) as a means of integrating the systems of all courts in the country, while prioritizing the development of the Electronic Judicial Process (PJe).

In the absence of standardized procedural platforms, a wide range of case management systems emerged, including Eletronic Case Management System (*Processo Judicial Eletrônico*, PJe), e-Proc, Projudi, Themis, and e-SAJ, in addition to proprietary platforms developed by individual courts. It is worth noting that there was, in fact, an attempt to standardize PJe as a single nationwide case management system. Indeed, CNJ Resolution No. 185/2013[10] established PJe as the national system for judicial proceedings (Art. 1), and Articles 34 et seq. set out deadlines for its implementation, with full deployment scheduled for 2018. By the end of this period, however, courts politically resisted the definitive implementation of PJe, based on different justifications, ranging from its technical shortcomings to the potential loss of investments already made in their own systems.[11] This resistance persisted over time,

---

7 BRAZIL. National Justice Council. *Resolução nº 65, de 16 de dezembro de 2008*. Dispõe sobre a uniformização do número dos processos nos órgãos do Poder Judiciário e dá outras providências. Brasília, 2008. Available at: https://atos.cnj.jus.br/atos/detalhar/119. Accessed on: Oct. 24, 2025.

8 BRAZIL. National Justice Council. *Resolução nº 46, de 18 de dezembro de 2007*. Cria as Tabelas Processuais Unificadas do Poder Judiciário e dá outras providências. Brasília, 2007. Available at: https://atos.cnj.jus.br/atos/detalhar/167. Accessed on: Oct. 24, 2025.

9 BRAZIL. National Justice Council. *Resolução nº 335, de 29 de setembro de 2020*. Institui política pública para a governança e a gestão de processo judicial eletrônico. Integra os tribunais do país com a criação da Plataforma Digital do Poder Judiciário Brasileiro - PDPJ-Br. Mantém o sistema PJe como sistema de Processo Eletrônico prioritário do Conselho Nacional de Justiça. Brasília, 2020f. Available at: https://atos.cnj.jus.br/atos/detalhar/3496. Accessed on: Oct. 24, 2025.

10 BRAZIL. National Justice Council. *Resolução nº 185, de 18 de dezembro de 2013*. Institui o Sistema Processo Judicial Eletrônico PJe como sistema de processamento de informações e prática de atos processuais e estabelece os parâmetros para sua implementação e funcionamento. Brasília, 2013. Available at: https://atos.cnj.jus.br/atos/detalhar/1933. Accessed on: Oct. 24, 2025.

11 MIGALHAS. *PJe ou e-Proc? Tribunais contestam resolução do CNJ sobre suspensão imediata de e-Proc*. Migalhas Quentes, 31 Out. 2019. Available at: https://www.migalhas.com.br/quen-

ultimately leading to the adoption of an interoperability model of the Brazilian Judiciary Digital Platform (PDPJ-Br).

3. TECHNICAL FEATURES OF CODEX

It is within this context of integration, automation, and the exponential growth in the volume of information produced that Codex emerges, serving CNJ Resolution No. 446, of March 14, 2022, as its primary legal basis.[12] In this resolution, CNJ established Codex as the official tool for extracting data (both structured and unstructured) from electronic judicial proceedings in progress within the National Judiciary, with the exception of data from the Supreme Court (*Supremo Tribunal Federal*) and from the CNJ itself.[13]

In turn, CNJ Ordinance No. 183, of June 2, 2022[14], provided that, excluding the Supreme Court, all courts were required to integrate their judicial case management systems into Codex by June 30, 2022, with further exceptions for administrative systems, systems for monitoring non-electronic proceedings, and systems for monitoring electronic proceedings that were to be fully deactivated by that same date. Finally, Article 7 of CNJ Resolution No. 446/2022[15] established the automated sub-

tes/314284/pje-ou-e-proc--tribunais-contestam-resolucao-do-cnj-sobre-suspensao-imediata-de-e-proc. Accessed on: Oct. 24, 2025.

12 BRAZIL. National Justice Council. *Resolução nº 446, de 14 de março de 2022*. Institui a plataforma Codex como ferramenta oficial de extração de dados estruturados e não estruturados dos processos judiciais eletrônicos em tramitação no Poder Judiciário Nacional e dá outras providências. Brasília, 2022e. Available at: https://atos.cnj.jus.br/atos/detalhar/4417. Accessed on: Oct. 24, 2025.

13 The exception is provided for in Article 1 of CNJ Resolution No. 446, which mandates implementation only for the courts listed in items II to VII of Article 92 of the Brazilian Federal Constitution. CNJ exclusion is justified, as its proceedings are predominantly administrative rather than strictly judicial.

14 BRAZIL. National Justice Council. *Portaria nº 183, de 2 de junho de 2022*. Fixa prazo para a integração dos sistemas judiciais eletrônicos em funcionamento nos Tribunais à Plataforma Codex. Brasília, 2022c. Available at: https://atos.cnj.jus.br/atos/detalhar/4570. Accessed on: Oct. 24, 2025.

15 BRAZIL. National Justice Council. *Resolução nº 446, de 14 de março de 2022*. Institui a platafor-

mission of information from the courts' databases.

Currently, the implementation of Codex is used as a performance assessment criterion for courts (CNJ Ordinance No. 411/2024[16]), and this criterion has been included in such evaluations since 2022 (see CNJ Ordinance No. 170/2022[17]; Ordinance No. 82/2023[18]; Ordinance No. 353/2023[19]). These criteria effectively incentivize courts to integrate with the platform by increasing their performance scores for statistical purposes.

With respect to its algorithmic structure, the platform's source code is not publicly available. Nevertheless, its architecture can be partially inferred, both from its characterization as a data lake and from information sporadically disclosed by the courts.

According to CNJ Resolution No. 446/2022[20], each court implements the tool for extracting structured and unstructured data within its own infrastructure (Art. 3, § 1). As suggested by Article 1, unstructured data refer to texts (potentially in their entirety), images, spreadsheets, and similar data. Structured data, in turn, refer to

ma Codex como ferramenta oficial de extração de dados estruturados e não estruturados dos processos judiciais eletrônicos em tramitação no Poder Judiciário Nacional e dá outras providências. Brasília, 2022e. Available at: https://atos.cnj.jus.br/atos/detalhar/4417. Accessed on: Oct. 24, 2025.

16 BRAZIL. National Justice Council. *Portaria nº 411, de 2 de dezembro de 2024*. Institui o Regulamento do Prêmio CNJ de Qualidade, ano 2025. Brasília, 2024c. Available at: https://atos.cnj.jus.br/atos/detalhar/5880. Accessed on: Oct. 24, 2025.

17 BRAZIL. National Justice Council. *Portaria nº 170, de 20 de maio de 2022*. Brasília, 2022b. Available at: https://atos.cnj.jus.br/atos/detalhar/4549. Accessed on: Oct. 24, 2025.

18 BRAZIL. National Justice Council. *Portaria nº 82, de 31 de março de 2023*. Institui o regulamento do Prêmio CNJ de Qualidade, ano 2023. Brasília, 2023a. Available at: https://atos.cnj.jus.br/atos/detalhar/5019. Accessed on: Oct. 24, 2025.

19 BRAZIL. National Justice Council. *Portaria nº 353, de 4 de dezembro de 2023*. Institui o Regulamento do Prêmio CNJ de Qualidade, ano 2024. Brasília, 2023b. Available at: https://atos.cnj.jus.br/atos/detalhar/5366. Accessed on: Oct. 24, 2025.

20 BRAZIL. National Justice Council. *Resolução nº 446, de 14 de março de 2022*. Institui a plataforma Codex como ferramenta oficial de extração de dados estruturados e não estruturados dos processos judiciais eletrônicos em tramitação no Poder Judiciário Nacional e dá outras providências. Brasília, 2022e. Available at: https://atos.cnj.jus.br/atos/detalhar/4417. Accessed on: Oct. 24, 2025.

metadata (procedural class, parties, adjudicating body, subject matter according to the TPU), mirroring the procedural metadata of DATAJUD, as provided in Article 6 of CNJ Resolution No. 331/2020[21]. These data are certainly converted into a standardized format (commonly JSON in this type of application) and subsequently tokenized, given the volume of data, for processing optimization. This is corroborated by a study conducted by the *Fundação Getulio Vargas* (FGV), which relied on data provided by the CNJ. The study expressly states that the entire process indeed involves the use of JSON.[22]

It is, therefore, a standard ETL (extract, transform, load) process. Regarding ETL procedures, Bolasco et al. (2005, p. 39) characterize them as follows[23]:

> *"Extraction Transformation Loading are aimed at filing non-structured textual material into categories and structured fields. The search engines are usually associated with ETL that guarantee the retrieval of information, generally by systems foreseeing conceptual browsing and questioning in natural language."*

Once the data are loaded into the central platform, data mining techniques can be applied, including for subsequent use in training of machine learning models and related applications, an aspect that will be analyzed below.

Due to the lack of transparency, considerations regarding the technical functioning presented above are based on practices observed in similar databases. The actual operation of Codex, however, is not known, which raises concerns regarding the collection, processing, and use of data, issues that will be examined in specific sec-

---

[21] BRAZIL. National Justice Council. *Resolução nº 331, de 20 de agosto de 2020*. Institui a Base Nacional de Dados do Poder Judiciário - DataJud como fonte primária de dados do Sistema de Estatística do Poder Judiciário - SIESPJ para os tribunais indicados nos incisos II a VII do art. 92 da Constituição Federal. Brasília, 2020e. Available at: https://atos.cnj.jus.br/atos/detalhar/3428. Accessed on: Oct. 24, 2025.

[22] FUNDAÇÃO GETULIO VARGAS. *Inteligência artificial: tecnologia aplicada à gestão dos conflitos no âmbito do Poder Judiciário brasileiro*. Rio de Janeiro: FGV Justiça, Out. 2024. Available at: https://justica.fgv.br/sites/default/files/2024-10/ia_2fase_0.pdf. Accessed on: Oct. 24, 2025.

[23] BOLASCO, S. et al. Understanding Text Mining: A Pragmatic Approach. *In*: SIRMAKESSIS, S. (ed.). *Knowledge Mining*: Proceedings of the NEMIS 2004 Final Conference (Studies in Fuzziness and Soft Computing). [S. l.]: Springer, 2005. p. 31-50.

tions of this chapter.

4. DISCUSSION ON THE POTENTIAL IMPACTS OF CODEX ON IMPROVING INFORMATION RETRIEVAL

Having set out the historical background, legal framework, and algorithmic structure of Codex, a judicial data lake is not only necessary, but indispensable in the context of Brazil's mass litigation. As previously demonstrated by Morais (2021)[24] and Morais (2023)[25], access to case-related information constitutes a pragmatic prerequisite for the functioning of a system of precedents. It is not possible to structure such a system if precedents themselves are not, first and foremost, accessible.

A robust database, on the scale already achieved by Codex, is the base from which a wide range of application possibilities emerges. Unique and structural opportunities for document data management in the Brazilian context are emphasized in subsequent sections. These proposals do not focus strictly on the computational techniques employed, which may vary according to the data analysis strategies adopted by each information manager.

4.1. Unified National Indexing

The immediate consequence of Codex is self-referential: Codex is a database that contributes to the improvement of search tools for retrieving its own data. Its vast repository serves to develop machine learning training models. This is evidenced by the existence of a specific tool developed by the CNJ, that is, JUMP-CNJ, for mining Codex data, as noted in the 2022 management transition report.[26]

---

[24] MORAIS, G. R. de. Why natural language processing is indispensable to entrech Brazil's precedents system. *In*: GOMES, A. de O. et al. (ed.). *Encontro de Administração da Justiça*: Anais do ENAJUS 2021. [S. l.]: IBEPES, 2021.

[25] MORAIS, G. R. de. *Pressupostos e diretrizes para a aplicação do processamento de linguagem natural à recuperação de informações jurisprudenciais*: uma aplicação da AI4SG. [S. l.]: Editora Dialética, 2023.

[26] BRAZIL. National Justice Council. *Relatório de transição SG v. 5: 2022 09 05*. Brasília, 2022d. Available at: https://bibliotecadigital.cnj.jus.br/jspui/bitstream/123456789/768/1/relatorio-de-transicao-sg-v5-2022-09-05.pdf. Accessed on: Oct. 24, 2025.

In legal documents, the primary objective of natural language processing (NLP) models concerns the optimization of information search and retrieval techniques.[27] Indeed, the strategic advantage of countries, such as China in the field of artificial intelligence, lies not primarily in a highly qualified workforce, but rather in the immense volume of data available for training learning models.[28] In this regard, the fact that Brazil possesses a robust and unified database, specifically designed for the legal sector, provides an ideal environment for the application of natural language processing techniques. The increase in the volume of information available for analysis is a fundamental factor in the development of NLP.[29]

On the other hand, the volume of information alone cannot be regarded as the central element. A larger volume of data does not necessarily result in higher-quality analysis. Incorrectly classified input data will likely produce undesirable outputs, as pointed in Datatilsynet (2018, p. 11)[30] and Nunes and Marques (2019, pp. 48-50).[31] In the context of large language models, it is well established that the quantity of information does not necessarily yield improvements in recall and precision in information retrieval. Nevertheless, the systemic advancement of artificial intelligence enables the development of incremental information retrieval models. Therefore, the use of such models has become a pragmatic imperative.

Regarding the unification of databases, while examining the problem of accessibility to case law data in Europe, Filtz, Kirrane, and Polleres (2018)[32] had pointed to

---

27 BOLASCO, S. et al. Understanding Text Mining: A Pragmatic Approach. *In*: SIRMAKESSIS, S. (ed.). *Knowledge Mining*: Proceedings of the NEMIS 2004 Final Conference (Studies in Fuzziness and Soft Computing). [S. l.]: Springer, 2005. p. 31-50.

28 FILIPOVA, I. A. Legal regulation of artificial intelligence: Experience of China. *Journal of Digital Technologies and Law*, [S. l.], v. 2, n. 1, p. 46-73, 2024.

29 KATZ, D. M. et al. *Natural language processing in the legal domain*. [S. l.]: arXiv, 23 Feb. 2023. Preprint. Available at: https://arxiv.org/abs/2302.12039. Accessed on: Oct. 24, 2025.

30 DATATILSYNET. *Artificial intelligence and privacy*. [S. l.], 2018.

31 NUNES, D.; MARQUES, A. L. P. C. Algoritmo: o risco da decisão das máquinas. *Bonijuris*, [S. l.], v. 31, n. 659, p. 44-58, 2019.

32 FILTZ, E. et al. Interlinking Legal Data. *In*: INTERNATIONAL CONFERENCE ON SEMANTIC SYSTEMS, 14th., 2018, [S. l.]. *Proceedings...* [S. l.]: CEUR-WS.org, 2018. p. 1-4. Available at: https://www.researchgate.net/publication/344378847_Interlinking_Legal_Data. Accessed on: Oct. 24, 2025.

the need for a unified database of European judicial proceedings. This phenomenon is commonly known as legal open data. Although the European judicial system presents complicating factors that do not exist in Brazil, such as the multiplicity of languages, legal traditions, and normative frameworks, the issue of multiple systems for storing case law information is similar in both legal systems. Among the factors that complicate jurisprudential analysis in European States is the need to create parsers capable of structuring legal documents while considering the linguistic and structural differences among them. Therefore, with respect to the unification of databases, the mere existence of Codex already represents an important advancement for Brazilian document management, capable of fostering the popularization of natural language processing in the Portuguese language, particularly due to the scale of its informational repository.

Regarding the analysis of information itself, there is a range of factors which, whether considered individually or cumulatively, may significantly enhance the efficiency of data mining and the advancement of indexing techniques.

Indexing is essentially a means of summarizing and representing the content of documents (Lancaster, 2004, pp. 3–6;[33] Savoy; Gaussier, 2010, p. 456)[34]. The role of the indexer, in turn, is to anticipate the needs of a potential user and to provide the means by which those needs can be met, calibrating the levels of recall and precision in the analysis (Lancaster, 2004, p. 17)[35]. Put simply, recall refers to the ability to retrieve relevant documents from the total universe of documents that could be retrieved; precision, by contrast, is a negative concept, relating to the non-retrieval of irrelevant documents (Morais, 2023, p. 116).[36] The correlation between these two measures tends to be inversely proportional: improvements in recall typically come at the expense of precision, and vice versa; moreover, the larger the database, the less

---

33 LANCASTER, F. W. *Indexação e resumos: teoria e prática*. 2ª ed. [S. l.]: Briquet de Lemos, 2004.

34 SAVOY, J.; GAUSSIER, E. Information retrieval. *In*: INDURKHYA, N.; DAMERAU, F. J. (ed.). *Natural Language Processing*. 2nd ed. [S. l.]: Chapman & Hall/CRC, 2010. p. 455-484.

35 LANCASTER, F. W. *Indexação e resumos: teoria e prática*. 2ª ed. [S. l.]: Briquet de Lemos, 2004.

36 MORAIS, G. R. de. *Pressupostos e diretrizes para a aplicação do processamento de linguagem natural à recuperação de informações jurisprudenciais*: uma aplicação da AI4SG. [S. l.]: Editora Dialética, 2023.

tolerable low precision levels becomes (Lancaster, 2003, p. 4).[37] This last observation is particularly applicable to Codex, given the breadth of its database.

An efficient indexing is the one capable of adjusting precision and recall rates according to its intended purpose. In addition, other factors also contribute to indexing efficiency, especially the delimitation of points of interest within a document, which requires the indexer to understand the needs of potential users (Lancaster, 2004, p. 15–17, 24).[38] In other words, each indexing process demands its own analytical criteria, depending on the purpose of the information retrieval being carried out.

Having made these initial considerations regarding the fundamentals of indexing, Brazilian jurisprudential big data already contains facilitating elements for its application within Codex itself. The first facilitating element is the existence of a foundational layer of thematic indexing, represented by the terms contained in the Unified Procedural Table (TPU). The further development of these entries, already provided in the TPU, reduces the time required for grouping information and for improvements in recall and precision rates. As discussed by Puente and Coalla (2023, p. 113) [39], in the field of textual analysis, the larger the corpus (database), the lower the occurrence of new items. The authors justify this correlation by noting that tokens (understood broadly as units of analysis) increase linearly with the size of the database, while token types tend to repeat themselves.[40] Lancaster (2004)[41] also identifies several techniques capable of contributing to the improvement of this foundational indexing layer, such as conceptual analysis,[42] translation,[43] controlled vocabulary

---

37 LANCASTER, F. W. *Indexação e resumos: teoria e prática*. 2ª ed. [S. l.]: Briquet de Lemos, 2004.

38 LANCASTER, F. W. *Indexação e resumos: teoria e prática*. 2ª ed. [S. l.]: Briquet de Lemos, 2004.

39 PUENTE, M.; COALLA, L. *Mineração de Dados Textuais: Conceitos e Técnicas*. [S. l.]: Blucher, 2023.

40 The encoding of the Portuguese language is provided for in ISO-8859-1. According to Palmer (2010, p. 12), this encoding only represents one possible method of tokenizing the Portuguese language.

41 LANCASTER, F. W. *Indexação e resumos: teoria e prática*. 2ª ed. [S. l.]: Briquet de Lemos, 2004.

42 This involves defining the subject of the document, the reasons for its incorporation into the collection, and the points of interest to the user within that document (Lancaster, 2004, p. 9).

43 Translation consists of converting a document into a set of indexing terms. This may be carried out either through extraction, in which indexing terms are obtained from terms already contained in the document, or through assignment, in which the terms are not originally present in the

development techniques,[44] the assignment of weights and standards to the vocabulary obtained,[45] and the adaptation of these techniques to automated extraction.

The advancement of these techniques, which are merely illustrative given the constant possibility of creating new methods, contributes to the construction of information retrieval systems oriented toward the needs of the intended user. Codex's pre-structured database makes the application of these techniques technically feasible, and their effectiveness may be further enhanced through the use of artificial intelligence-based models.

The standardization of judicial decision structures consists in another facilitating element for the indexing of Brazilian judicial big data. The standardization of textual structure enables document analysis algorithms to be programmed in a manner that optimizes their analysis on sections that contains the greatest documentary relevance. This constitutes the first theoretical guideline for the development of natural language processing in law, as proposed by Morais (2023, pp. 121–126).[46]

Still at an incipient stage, Recommendation No. 154/2024 of the National Council of Justice[47] establishes a standardized model for judicial headnotes. Indeed, there is

document (Lancaster, 2004, p. 18).

44 A controlled vocabulary is, essentially, a dictionary of terms relevant to the analysis. It may relate either to the substantive content of the text, that is, the words relevant for identifying the subject matter of the document, or to the formal content of the text, namely the words that compose its syntactic structure. In this regard, see Lancaster (2004, p. 19) and Savoy and Gaussier (2010, p. 460).

45 Weight assignment is one among several techniques that may be used to overcome the (semantic) problem of polysemy, as well as the issue of indexing exhaustivity, since if a single word with greater weight is capable of adequately indexing the content of a text, the need for additional words in the index is eliminated. In this regard, see Câmara Júnior (2013, p. 90); Savoy and Gaussier (2010, p. 458), who even propose combining qualitative weighting factors related to the efficiency of a term as an indexer with quantitative factors related to the frequency of term occurrence; and Lancaster (2004, p. 186). For decades, weight assignment has been regarded as a promising method for improving document classification, including the use of AI-based automated indexers, as already demonstrated by Jenkins in 2008 (p. 598).

46 MORAIS, G. R. de. *Pressupostos e diretrizes para a aplicação do processamento de linguagem natural à recuperação de informações jurisprudenciais*: uma aplicação da AI4SG. [S. l.]: Editora Dialética, 2023.

47 BRAZIL. National Justice Council. *Recomendação nº 154, de 13 de agosto de 2024*. Recomenda a

a broader movement toward data centralization and standardization that is not limited to storage architecture itself (namely, the Codex platform) or to application hubs such as the BDPJ. The CNJ's regulatory centralization tends to unify document formatting itself. In fact, standardization constitutes a positive aspect for large-scale data analysis as it accelerates automated processing. It is unnecessary for extraction and indexing mechanisms to map document structures from scratch, thereby reducing processing time.[48] Thus, one of the pre-processing stages is partially eliminated, namely document normalization, understood as the conversion of a set of digital files into a standardized formatted document (Câmara Júnior, 2013, p. 38).[49]

An analysis of National Council of Justice Recommendation No. 154/2024[50] reveals a normative trend toward the standardization of macro-textual elements, particularly judicial syllabi. The progression toward the uniformization of micro-textual elements, however, is considerably more complex, if not impossible. Such an effort would seek to standardize the very structure through which ideas are expressed within smaller textual units (paragraphs and sentences), while also eliminating the writer's linguistic biases. This type of uniformization would suppress the writer's distinctive characteristics, which is neither feasible nor desirable in adjudicative activity itself.

---

todos os tribunais do país a adoção de modelo padronizado de elaboração de ementas (ementa-padrão). Brasília, 2024d. Available at: https://atos.cnj.jus.br/atos/detalhar/5693. Accessed on: Oct. 24, 2025.

48 It should be noted that the identification of textual structures constitutes part of the pre-processing stage (Nastase, Mihalcea, and Radev, 2015, p. 670). Although the mapping of textual structures may appear trivial, the first stage of Project Victor consisted precisely in classifying procedural documents according to the type of petition filed. Not only did this produce substantial efficiency gains for the sectors responsible for such classification, but it also served as a foundational element for structuring the subsequent stages of the project's development (Silva et al., 2018, p. 1).

49 Câmara Júnior, A. T. da. (2013). Processamento de linguagem natural para indexação automática semântico-ontológica [Doctoral thesis, University of Brasília].

50 BRAZIL. National Justice Council. *Recomendação nº 154, de 13 de agosto de 2024*. Recomenda a todos os tribunais do país a adoção de modelo padronizado de elaboração de ementas (ementa-padrão). Brasília, 2024d. Available at: https://atos.cnj.jus.br/atos/detalhar/5693. Accessed on: Oct. 24, 2025.

These two facilitating factors, namely, a pre-established thematic indexing and the emergence of policies aimed at the standardization of macro-textual elements, combined with the robustness of the Codex Platform enable the development of a unified nationwide indexing system specifically designed for the legal field. Its implementation may improve the balance between precision and recall rates, thereby optimizing the outcome expected by the end user, specially the retrieval of information effectively sought.

4.2. Nationwide Semantic Retrieval

The second operational advantage of the Codex Platform concerns the semantic analysis of judicial texts. Originally, information retrieval techniques were focused on specialized fields of knowledge. Each field of knowledge required its own specific retrieval approach. However, with the advent of the internet, information retrieval began to be performed across multiple documents simultaneously through multi-document and open-domain analysis (Hobbs; Riloff, 2010, pp. 514–515).[51] However, non-specialized analysis proves problematic for semantic understanding, as it is inherently contextual. Decades ago, it had already been observed that, for purposes of thematic analysis, text mining must invariably take into account the context in which a document is inserted (Tan, 1999, p. 5).[52]

In fact, the ultimate objective of an indexing tool is to understand the meaning of a word in order to classify it accurately, a meaning that effectively varies according to context (Nastase, Mihalcea, Radev, 2015, p. 681).[53] A database on the scale of Codex may provide the necessary foundation for the development of a semantic dictionary specifically adapted to the legal domain.

---

[51] HOBBS, J. R.; RILOFF, E. Information extraction. *In*: INDURKHYA, N.; DAMERAU, F. J. (ed.). *Natural Language Processing*. 2nd ed. [S. l.]: Chapman & Hall/CRC, 2010. p. 511-532.

[52] TAN, A.-H. Text mining: The state of the art and the challenges. *In*: WORKSHOP ON KNOWLEDGE DISCOVERY FROM ADVANCED DATABASES, 1999. *Annals...* [S. l.]: [s. n.], 1999. p. 65-70.

[53] NASTASE, V.; MIHALCEA, R.; RADEV, D. R. A survey of graphs in natural language processing. *Natural Language Engineering*, [S. l.], v. 21, n. 5, p. 665-698, 2015.

As noted by Yarowsky (2010, p. 315)[54], disambiguation is, essentially, a classification problem. One of the ways to achieve robust and efficient classification is through the creation of a semantic inventory (Hobbs; Riloff, 2010, p. 526).[55] Such a dictionary functions as an enhancement of the indexing structure discussed in the previous subsection, with a significant impact on the information retrieval efficiency. Its construction may rely on techniques such as thesaurus development (bag-of-words)[56], tagging[57], concept hierarchies[58], weight assignment[59], among many others.

Regardless of the technique employed, the creation of a semantic dictionary specifically designed to the analysis of judicial decisions, such as the development of the unified national indexer, constitutes an advanced application of natural language

---

54 YAROWSKY, D. Word sense disambiguation. *In*: INDURKHYA, N.; DAMERAU, F. J. (ed.). *Natural Language Processing*. 2nd ed. [S. l.]: Chapman & Hall/CRC, 2010. p. 315-338.

55 Hobbs, J. R., & Riloff, E. (2010). Information extraction. In N. Indurkhya & F. J. Damerau (Eds.), *Natural Language Processing* (2nd ed., pp. 511–532). Chapman & Hall/CRC.

56 This technique consists of an inventory of words associated with a specific topic. The topic is subsequently divided into subtopics containing words specific to each subclassification. The number of subclassifications varies according to the level of granularity intended by the indexer (Yarowsky, 2010, p. 318). Some authors, such as Winkels (2018, p. 72), describe this method as excessively simplistic for achieving satisfactory results. Nevertheless, the simplicity of the method alone is not a sufficient reason to dismiss it, given that the required level of accuracy varies according to the purpose of the indexing process, as previously discussed in this paper.

57 Tagging consists of associating a thematic attribute with a text through a classification label. It is generally employed as a complementary technique alongside other forms of indexing. For instance, tagging is commonly used in thesaurus-based indexing systems.

58 Technique consisting in the creation of hierarchical structures for organizing words by identifying the meaning of a term from the most specific to the most general level. Yarowsky (2010, p. 317) illustrates this approach using the example of assigning meaning to the word *crane* (which may refer either to a lifting machine or to a species of bird). In English, the distinct semantic senses would correspond to distinct hierarchies, such as: *crane* (lifting machine) > inanimate object > physical entity; and *crane* (bird) > living being > physical entity. Both converge at a more generalized level (*physical entity*), but differ at an intermediate level, as one denotes a living being whereas the other does not.

59 Weight assignment, much like tagging, is generally employed as a complementary technique. Depending on the context in which a word appears, greater weight is assigned to its most probable meaning. This technique may even be used as an attribute associated with the labels employed in tagging systems.

processing potentially enabled by the use of Codex. The development of both techniques has the potential to substantially improve the retrieval of Brazilian case law information.

5. REGULATORY CHALLENGES RELATED TO THE IMPLEMENTATION OF CODEX AND ITS INTEGRATION WITH AI MODELS

Once the operational advantages offered by Codex have been presented, it is appropriate to examine the challenges associated with its implementation, as well as the regulation of natural language processing models that rely on this type of massive and partially structured database , which potentially contains sensitive data.

Among the elements requiring contextualized and sector-specific regulatory attention for the development of natural language processing based on Codex data are issues such as data access, information storage security (information security), and the protection of the fundamental right to personal data protection.

5.1. Data Access and APIs

Access to Codex data involves three distinct issues: who may access the data, on what legal basis such access is granted, and through which mechanisms or interfaces such access is implemented.

As a constitutional imperative concerning procedural acts, publicity is presumed to be the rule, whereas confidentiality constitutes an exception (Brazilian Federal Constitution, art. 5, LX, and art. 93, IX). As a constitutional guarantee, the publicity of procedural acts constitutes an entrenched constitutional clause under article 60, § 4, IV of Brazilian Federal Constitution. Article 189 of Brazil's Code of Civil Procedure (CPC) operates within the same constitutional framework, insofar as the imposition of judicial secrecy, whether concerning specific acts or the proceedings as a whole, is regarded as highly exceptional. This infra-constitutional provision also applies, by extension, to electoral, labor, and administrative proceedings due to the supplementary and subsidiary application of this rules, as provided in article 15 of the CPC. Accordingly, it may be presumed that most of the data contained in Codex is public,

whereas only a smaller portion is subject to confidentiality, given the correlation between the principle of publicity and the exceptional nature of secrecy.

As a fundamental constitutional guarantee, any Brazilian citizen possesses a subjective right to access procedural data stored within the Codex system, except for confidential information. This principle aligns with CNJ Resolution No. 121/2010[60], which restricts public access to certain procedural documents, a practice that remains constitutionally questionable. Nevertheless, the specific means of accessing such data must be clarified.

To date, the only publicly available Codex data are limited to the monitoring dashboard referenced in footnote 1. This reflects purely statistical information rather than direct data access. However, CNJ Resolution No. 574/2024[61] indicates that an Application Programming Interface (API) will likely be provided to access the Codex repository, subject to a specific financial or non-financial reciprocal obligation. Concurrently, Ordinance No. 316/2024[62] regulates the technical requirements to enable this access, including compliance with security protocols.

Enabling APIs appears to be the primary objective of the Codex system, as data lakes of this magnitude are expected to support massive data access. To effectively implement open justice principles while safeguarding data protection, APIs would provide a specific access mechanism. This approach would reduce the burden of web scrapers on other court systems while ensuring efficient data access for subsequent processing, mining, parameter correlation, and other jurimetrics applications.

---

60 BRAZIL. National Justice Council. *Resolução nº 121, de 5 de outubro de 2010*. Dispõe sobre a divulgação de dados processuais eletrônicos na rede mundial de computadores, expedição de certidões judiciais e dá outras providências. Brasília, 2010. Available at: https://atos.cnj.jus.br/atos/detalhar/92. Accessed on: Oct. 24, 2025.

61 BRAZIL. National Justice Council. *Resolução nº 574, de 26 de agosto de 2024*. Dispõe sobre o acesso a dados judiciais públicos consolidados pelo Conselho Nacional de Justiça, prevê a possibilidade de depósito de serviços privados na PDPJ-Br e institui o portal unificado para usuários internos. Brasília, 2024a. Available at: https://atos.cnj.jus.br/atos/detalhar/5708. Accessed on: Oct. 24, 2025.

62 BRAZIL. National Justice Council. Portaria nº 316, de 20 de setembro de 2024. Regulamenta o acesso a dados judiciais públicos consolidados pelo Conselho Nacional de Justiça. Brasília, 2024b. Available at: https://atos.cnj.jus.br/atos/detalhar/5792. Accessed on: Oct. 24, 2025.

Although this API access is neither fully regulated nor publicly available, society at large is expected to gain access to the Codex system for the macro-analysis of public procedural data.

5.2. Information Security

According to Saltzer and Schroeder's classic triad,[63] information breaches fall into three categories: unauthorized data leakage, unauthorized information modification, and unauthorized denial of access. Within the Codex framework, these vulnerabilities analogously correspond to unauthorized data breaches of legally protected confidential data, the modification of procedural documents, and the unlawful denial of user access.

Unauthorized data leakage is the most evident challenge associated with Codex. This issue arises because a core principle of information security is the distribution of data across separate repositories. Such an approach reflects the informational segregation principle known as *the least common mechanism*, one of the eight security principles proposed by Saltzer and Schroeder (1975).[64] According to this principle, when a single shared component or mechanism is used by every user, compromising that one element is enough to create a relevant point of failure that exposes all data in the repository.

The second type of informational vulnerability, involving data manipulation or alteration, recalls the recent ransomware incident affecting the case repository of the *Superior Tribunal de Justiça* (Brazilian Superior Court of Justice).[65] On November 3, 2020, the court's case management system became the target of a cyberattack in

---

63 SALTZER, J.; SCHROEDER, M. *The protection of information in computer systems*. [S. l.]: University of Virginia, 1975. Available at: https://www.cs.virginia.edu/~evans/cs551/saltzer/. Accessed on: Oct. 24, 2025. p. 5

64 SALTZER, J.; SCHROEDER, M. *The protection of information in computer systems*. [S. l.]: University of Virginia, 1975. Available at: https://www.cs.virginia.edu/~evans/cs551/saltzer/. Accessed on: Oct. 24, 2025. p. 5

65 BRAZIL. Superior Tribunal de Justiça. *Comunicado da Presidência do STJ*. Brasília, 19 Nov. 2020. Available at: https://www.stj.jus.br/sites/portalp/Paginas/Comunicacao/Noticias/19112020-Comunicado-da-Presidência-do-STJ.aspx. Accessed on: Oct. 24, 2025.

which intruders blocked institutional access to judicial records, encrypted the stored data, and demanded payment for its recovery. To date, no suspects have been formally charged, and the compromised data were not recovered. The solution adopted by the court relied on restoring information from backup copies. This episode is illustrative as, despite significant advances in information security within large databases, particularly those maintained by government institutions, vulnerabilities *always* remain. Information security ultimately consists of identifying such weaknesses and addressing or mitigating them, especially through redundancy strategies.

One week after the cyberattack against the Brazilian Superior Court of Justice, on November 10, 2020, the Brazilian National Council of Justice issued Ordinance No. 242/2020 establishing the Judiciary Cybersecurity Committee. Its purpose was to coordinate the development of four key strategic documents: the Cyber Incident Prevention Protocol[66]; the Cyber Crisis Management Protocol[67]; the Protocol for the Investigation of Cyber Incidents[68]; and the Judiciary Cybersecurity and Information Security Strategy[69]. Currently, CNJ Ordinance No. 162/2021 consolidates the three information security protocols within its annexes.

The first protocol, namely the Cyber Incident Prevention Protocol, is set out in Annex I of CNJ Ordinance No. 162/2021. Its primary focus lies in establishing guidelines for training teams responsible for cybersecurity within the Judiciary, as well as identifying and containing threats. Among its various guidelines, particular attention should be given to the recognition of a fundamental security practice: cybersecurity is understood as a *collective endeavor*.

The second protocol, namely the Cyber Crisis Management Protocol, is contained

---

66 Initially regulated by Ordinance No. 292 of December 17, 2020. This protocol was later repealed by Ordinance No. 162 of June 10, 2021, which approved the protocols established under Resolution No. 396 of June 7, 2021.

67 Initially regulated by Ordinance No. 290 of December 17, 2020. This protocol was later repealed by Ordinance No. 162 of June 10, 2021, which approved the protocols established under Resolution No. 396 of June 7, 2021.

68 Initially regulated by Ordinance No. 291 of December 17, 2020. This protocol was later repealed by Ordinance No. 162 of June 10, 2021, which approved the protocols established under Resolution No. 396 of June 7, 2021.

69 Established by Resolution No. 396 of June 7, 2021, which repealed the previous protocols.

in Annex II of CNJ Ordinance No. 162/2021. Unlike the first protocol, this document provides concrete and structured procedures for addressing pre-crisis, in-crisis, and post-crisis situations. During the pre-crisis stage, the protocol requires the identification of critical activities and the establishment of specific procedures. During a crisis, predefined protocols are implemented, with adjustments according to the nature of the event. In the post-crisis stage, measures include identifying security failures that contributed to the incident, preparing reports aimed at mitigating future crises, and establishing recovery measures.

The third and final protocol, namely the Protocol for the Investigation of Cyber Incidents, is contained in Annex III of CNJ Ordinance No. 162/2021. It establishes guidelines for preserving evidence related to authorship and materiality of cyberattacks, as well as communication procedures for reporting incidents to judicial police authorities and the Public Prosecutor's Office, particularly through a *Computer Network Security Incident Communication Report*.

The Judiciary Cybersecurity and Information Security Strategy is completed by Annexes IV to VIII of CNJ Ordinance No. 162/2021. Annexes IV to VI consist of reference manuals containing technical standards for infrastructure protection, threat prevention and mitigation, identity management, and access control. Annex VII provides a reference framework for cybersecurity education and awareness policies, while Annex VIII contains a glossary.

With specific regard to Codex, its opacity prevents external stakeholders from assessing whether its architecture complies with the National Cybersecurity Strategy for the Brazilian Judiciary. There is no structured dialogue involving the CNJ, universities, judicial police authorities, the Brazilian Bar Association, and other actors directly concerned with protecting the legal data stored within Codex. This highly closed model of information governance conflicts with security good practices established in Annex I (Cyber Incident Prevention Protocol) of CNJ Ordinance No. 162/2021, as well as with the transparency guidelines set out in Annex V of the same ordinance (Manual for Threat Prevention and Mitigation), particularly items 13.1(f) and 12.1(f). In other words, CNJ recognizes transparency as an important component of information risk management while simultaneously restricts such transparency by near-total opacity surrounding Codex.

Consequently, this opacity prevents existing vulnerabilities, inevitably present in greater or lesser extent, from being subjected to public scrutiny aimed at improving the system. While broader access to its architecture could, on the one hand, increase exposure to malicious attacks, greater openness could also enable a collaborative approach to system improvement. This, in fact, aligns with the concept of open design proposed by Saltzer and Schroeder (1975).[70]

Thus, from a data governance perspective, there are important gaps that require institutional improvements by the CNJ, particularly regarding Codex's information security policy.

5.3. Personal Data Protection

From an infrastructure perspective, information security protects the *container*, namely, the systems responsible for storing and processing data. From the perspective of the data itself, data protection safeguards the *content* being stored, that is, the data itself. In the case of personal data, such information forms part of the personality rights of natural persons, as derived from a systemic interpretation of Article 5, items X, XII, and LXXIX of the Federal Constitution; Articles 11 and following of the Civil Code; and Article 2, items I, IV, and VII of Federal Law No. 13.709/2018 (LGPD, that is, Brazilian GDPR), as recognized by the Federal Supreme Court in ADI No. 6.387/DF. By extension, such protection also applies to legal entities pursuant to Article 52 of the Civil Code.

With regard to the protection of data stored within Codex, Article 9 of CNJ Resolution No. 446/2022 makes direct reference to the Brazilian GDPR. This explicit reference removes any doubt regarding the applicability of the statute. Although such an express provision was not strictly necessary, Brazilian GDPR would apply in any event pursuant to Article 3, *caput*, and because none of the exceptions set forth in Article 4 of the same statute are applicable.

Thus, it is necessary to distinguish among the different categories of data con-

[70] SALTZER, J.; SCHROEDER, M. *The protection of information in computer systems*. [S. l.]: University of Virginia, 1975. Available at: https://www.cs.virginia.edu/~evans/cs551/saltzer/. Accessed on: Oct. 24, 2025.

tained in judicial proceedings prior to determining the legal treatment applicable to each of them. As a legal requirement, judicial proceedings begin with the submission of the personal data of both claimant and defendant, except where such information is unknown, in which case judicial measures must be taken to obtain it (Article 319, item II and §1 of Brazil's Code of Civil Procedure). Furthermore, depending on the nature of the dispute, court records may also contain sensitive personal data. This is the case, for example, in healthcare-related litigation, where the submission of medical records and examination results is often required.

An analysis of CNJ Resolution No. 446/2022 shows that Article 2, §3 broadly assigns to the CNJ responsibility for ensuring confidentiality, access control, and the definition of document secrecy levels. However, the technical procedures employed by the CNJ to achieve these objectives remain unknown. As discussed in the previous subsection, the opacity surrounding actual data management practices prevents individuals and institutions outside the CNJ from contributing to the identification of meaningful areas for improvement. More importantly, it also prevents any assessment of Codex's level of compliance with Brazilian GDPR itself.

6. FINAL REMARKS

Codex has emerged as an important repository of Brazilian judicial documents and may eventually become one of the largest of such repositories worldwide, given the context of large-scale litigation in which it operates.

Given the scale of its database, Codex has the potential to support the development of unified national indexing systems and nationwide semantic retrieval mechanisms adapted to the legal domain. These data-processing techniques could substantially improve the efficiency of legal information retrieval and jurisprudential analysis. Particular attention should also be given to data access governance through APIs, particularly in light of CNJ Ordinance No. 316/2024, which provides the mechanism capable of enabling the effective implementation of such techniques.

However, Codex's current opacity prevents a detailed assessment of key aspects of the platform. In summary: (a) there is no public transparency or auditability regarding its development, with very limited information available on its evolution over time; (b) there is no available information regarding its algorithmic architecture,

even at a general level; (c) there is no information security program guided by transparency and integration of different sectors of society, such as universities, judicial police authorities, the Brazilian Bar Association, and the Public Prosecutor's Office, among others. It is important that these shortcomings be addressed by the CNJ throughout the implementation and development of Codex, particularly to foster a more dialogical and collaborative engagement among different social actors.

Having examined these elements, it becomes clear that the regulation of natural language processing model development necessarily involves an assessment of the core characteristics of the databases that support such models. Accordingly, the fundamental regulatory issues surrounding Codex are, to a large extent, reflected in the NLP models applied to the platform. Central concerns include lawful forms of access to *public* judicial data, the purposes for which such data are analyzed, and the proper handling of data in accordance with the data protection framework currently in force in Brazil.

There is little doubt that natural language processing represents one of the most effective approaches for handling the analysis of Brazil's massive volume of judicial information, particularly in supporting the consolidation of the precedent system. The development of NLP systems, however, requires transparency, auditability, and the implementation of ethical principles in data governance. This necessarily includes adherence to such standards by the very repositories that provide the underlying data on which these systems depend.

Like the Library of Babel, a virtually infinite repository of information requires adequate retrieval mechanisms to enable efficient access to relevant knowledge. Such a library encompasses all existing knowledge; thus, its recall rate tends toward *infinity*. However, without mechanisms capable of refining search precision, its efficiency becomes meaningless, and information ultimately deteriorates into noise.

REFERENCES

BASILE, J. *Library of Babel*. [S. l.], [n.d.]. Available at: https://libraryofbabel.info. Accessed on: Oct. 24, 2025.

BOLASCO, S. et al. Understanding Text Mining: A Pragmatic Approach. *In*: SIRMAKESSIS, S. (ed.). *Knowledge Mining*: Proceedings of the NEMIS 2004 Final Conference (Studies in Fuzziness and

Soft Computing). [S. l.]: Springer, 2005. p. 31-50.

BORGES, J. L. A biblioteca de Babel. *In*: BORGES, J. L. *O jardim de veredas que se bifurcam*. [S. l.]: Editora Globo S.A., 1999.

BRAZIL. [Constituição (1988)]. *Constituição da República Federativa do Brasil, de 5 de outubro de 1988*. Brasília, 1988. Available at: http://www.planalto.gov.br/ccivil_03/constituicao/constituicao.htm. Accessed on: Oct. 24, 2025.

BRAZIL. *Lei nº 10.406, de 10 de janeiro de 2002*. Institui o Código Civil. Brasília, 2002. Available at: http://www.planalto.gov.br/ccivil_03/leis/2002/110406compilada.htm. Accessed on: Oct. 24, 2025.

BRAZIL. *Lei nº 13.105, de 16 de março de 2015*. Código de Processo Civil. Brasília, 2015. Available at: http://www.planalto.gov.br/ccivil_03/_ato2015-2018/2015/lei/113105.htm. Accessed on: Oct. 24, 2025.

BRAZIL. *Lei nº 13.709, de 14 de agosto de 2018*. Lei Geral de Proteção de Dados Pessoais (LGPD). Brasília, 2018. Available at: http://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/L13709compilado.htm. Accessed on: Oct. 24, 2025.

BRAZIL. National Justice Council. *Portaria nº 82, de 31 de março de 2023*. Institui o regulamento do Prêmio CNJ de Qualidade, ano 2023. Brasília, 2023a. Available at: https://atos.cnj.jus.br/atos/detalhar/5019. Accessed on: Oct. 24, 2025.

BRAZIL. National Justice Council. *Portaria nº 118, de 13 de abril de 2021*. Dispõe sobre o portfólio de soluções de tecnologia da informação e comunicação e serviços digitais do Conselho Nacional de Justiça. Brasília, 2021a. Available at: https://atos.cnj.jus.br/atos/detalhar/3866. Accessed on: Oct. 24, 2025.

BRAZIL. National Justice Council. *Portaria nº 162, de 10 de junho de 2021*. Aprova Protocolos e Manuais criados pela Resolução CNJ nº 396/2021, que instituiu a Estratégia Nacional de Segurança Cibernética do Poder Judiciário (ENSEC-PJ). Brasília, 2021b. Available at: https://atos.cnj.jus.br/atos/detalhar/3982. Accessed on: Oct. 24, 2025.

BRAZIL. National Justice Council. *Portaria nº 170, de 20 de maio de 2022*. Brasília, 2022b. Available at: https://atos.cnj.jus.br/atos/detalhar/4549. Accessed on: Oct. 24, 2025.

BRAZIL. National Justice Council. *Portaria nº 183, de 2 de junho de 2022*. Fixa prazo para a integração dos sistemas judiciais eletrônicos em funcionamento nos Tribunais à Plataforma Codex. Brasília, 2022c. Available at: https://atos.cnj.jus.br/atos/detalhar/4570. Accessed on: Oct. 24, 2025.

BRAZIL. National Justice Council. *Portaria nº 242, de 10 de novembro de 2020*. Institui o Comitê de Segurança Cibernética do Poder Judiciário. Brasília, 2020a. Available at: https://atos.cnj.jus.br/atos/detalhar/3566. Accessed on: Oct. 24, 2025.

BRAZIL. National Justice Council. *Portaria nº 290, de 17 de dezembro de 2020*. Institui o Protocolo de Gerenciamento de Crises Cibernéticas no âmbito do Poder Judiciário (PGCC/PJ). Brasília, 2020b.

Available at: https://atos.cnj.jus.br/atos/detalhar/3640. Accessed on: Oct. 24, 2025.

BRAZIL. National Justice Council. *Portaria nº 291, de 17 de dezembro de 2020*. Institui o Protocolo de Investigação para Ilícitos Cibernéticos no âmbito do Poder Judiciário. Brasília, 2020c. Available at: https://atos.cnj.jus.br/atos/detalhar/3641. Accessed on: Oct. 24, 2025.

BRAZIL. National Justice Council. *Portaria nº 292, de 17 de dezembro de 2020*. Determina a adoção de Protocolo de Prevenção a Incidentes Cibernéticos no âmbito do Poder Judiciário (PPICiber/PJ). Brasília, 2020d. Available at: https://atos.cnj.jus.br/atos/detalhar/3651. Accessed on: Oct. 24, 2025.

BRAZIL. National Justice Council. *Portaria nº 316, de 20 de setembro de 2024*. Regulamenta o acesso a dados judiciais públicos consolidados pelo Conselho Nacional de Justiça. Brasília, 2024b. Available at: https://atos.cnj.jus.br/atos/detalhar/5792. Accessed on: Oct. 24, 2025.

BRAZIL. National Justice Council. *Portaria nº 353, de 4 de dezembro de 2023*. Institui o Regulamento do Prêmio CNJ de Qualidade, ano 2024. Brasília, 2023b. Available at: https://atos.cnj.jus.br/atos/detalhar/5366. Accessed on: Oct. 24, 2025.

BRAZIL. National Justice Council. *Portaria nº 411, de 2 de dezembro de 2024*. Institui o Regulamento do Prêmio CNJ de Qualidade, ano 2025. Brasília, 2024c. Available at: https://atos.cnj.jus.br/atos/detalhar/5880. Accessed on: Oct. 24, 2025.

BRAZIL. National Justice Council. *Recomendação nº 154, de 13 de agosto de 2024*. Recomenda a todos os tribunais do país a adoção de modelo padronizado de elaboração de ementas (ementa-padrão). Brasília, 2024d. Available at: https://atos.cnj.jus.br/atos/detalhar/5693. Accessed on: Oct. 24, 2025.

BRAZIL. National Justice Council. *Relatório de transição SG v. 5: 2022 09 05*. Brasília, 2022d. Available at: https://bibliotecadigital.cnj.jus.br/jspui/bitstream/123456789/768/1/relatorio-de-transicao-sg-v5-2022-09-05.pdf. Accessed on: Oct. 24, 2025.

BRAZIL. National Justice Council. *Resolução nº 46, de 18 de dezembro de 2007*. Cria as Tabelas Processuais Unificadas do Poder Judiciário e dá outras providências. Brasília, 2007. Available at: https://atos.cnj.jus.br/atos/detalhar/167. Accessed on: Oct. 24, 2025.

BRAZIL. National Justice Council. *Resolução nº 65, de 16 de dezembro de 2008*. Dispõe sobre a uniformização do número dos processos nos órgãos do Poder Judiciário e dá outras providências. Brasília, 2008. Available at: https://atos.cnj.jus.br/atos/detalhar/119. Accessed on: Oct. 24, 2025.

BRAZIL. National Justice Council. *Resolução nº 121, de 5 de outubro de 2010*. Dispõe sobre a divulgação de dados processuais eletrônicos na rede mundial de computadores, expedição de certidões judiciais e dá outras providências. Brasília, 2010. Available at: https://atos.cnj.jus.br/atos/detalhar/92. Accessed on: Oct. 24, 2025.

BRAZIL. National Justice Council. *Resolução nº 185, de 18 de dezembro de 2013*. Institui o Sistema Processo Judicial Eletrônico PJe como sistema de processamento de informações e prática de atos

processuais e estabelece os parâmetros para sua implementação e funcionamento. Brasília, 2013. Available at: https://atos.cnj.jus.br/atos/detalhar/1933. Accessed on: Oct. 24, 2025.

BRAZIL. National Justice Council. *Resolução nº 331, de 20 de agosto de 2020*. Institui a Base Nacional de Dados do Poder Judiciário - DataJud como fonte primária de dados do Sistema de Estatística do Poder Judiciário - SIESPJ para os tribunais indicados nos incisos II a VII do art. 92 da Constituição Federal. Brasília, 2020e. Available at: https://atos.cnj.jus.br/atos/detalhar/3428. Accessed on: Oct. 24, 2025.

BRAZIL. National Justice Council. *Resolução nº 335, de 29 de setembro de 2020*. Institui política pública para a governança e a gestão de processo judicial eletrônico. Integra os tribunais do país com a criação da Plataforma Digital do Poder Judiciário Brasileiro - PDPJ-Br. Mantém o sistema PJe como sistema de Processo Eletrônico prioritário do Conselho Nacional de Justiça. Brasília, 2020f. Available at: https://atos.cnj.jus.br/atos/detalhar/3496. Accessed on: Oct. 24, 2025.

BRAZIL. National Justice Council. *Resolução nº 396, de 7 de junho de 2021*. Institui a Estratégia Nacional de Segurança Cibernética do Poder Judiciário (ENSEC-PJ). Brasília, 2021c. Available at: https://atos.cnj.jus.br/atos/detalhar/3975. Accessed on: Oct. 24, 2025.

BRAZIL. National Justice Council. *Resolução nº 446, de 14 de março de 2022*. Institui a plataforma Codex como ferramenta oficial de extração de dados estruturados e não estruturados dos processos judiciais eletrônicos em tramitação no Poder Judiciário Nacional e dá outras providências. Brasília, 2022e. Available at: https://atos.cnj.jus.br/atos/detalhar/4417. Accessed on: Oct. 24, 2025.

BRAZIL. National Justice Council. *Resolução nº 574, de 26 de agosto de 2024*. Dispõe sobre o acesso a dados judiciais públicos consolidados pelo Conselho Nacional de Justiça, prevê a possibilidade de depósito de serviços privados na PDPJ-Br e institui o portal unificado para usuários internos. Brasília, 2024a. Available at: https://atos.cnj.jus.br/atos/detalhar/5708. Accessed on: Oct. 24, 2025.

BRAZIL. National Justice Council. *Inteligência artificial: parceria com tribunal de Rondônia aproxima o futuro*. Brasília, 2018. Available at: https://www.cnj.jus.br/inteligencia-artificial-parceria-com-tribunal-de-rondonia-aproxima-o-futuro/. Accessed on: Oct. 24, 2025.

BRAZIL. National Justice Council. *Justice 4.0 Program*. Brasília, 2022a. Available at: https://www.cnj.jus.br/wp-content/uploads/2022/05/justice-4-0-program.pdf. Accessed on: Oct. 24, 2025.

BRAZIL. National Justice Council. *Painel de acompanhamento da Plataforma Codex*. Brasília, [n.d.-a]. Available at: https://paineisanalytics.cnj.jus.br/single/?appid=de543e38-1c79-412e-b7b2-51624b55349d&sheet=2c946c67-9efe-4e23-a74c-d1c4792fb0bb. Accessed on: Oct. 24, 2025.

BRAZIL. National Justice Council. *Plataforma Codex*. Brasília, [n.d.-b]. Available at: https://www.cnj.jus.br/sistemas/plataforma-codex/. Accessed on: Oct. 24, 2025.

BRAZIL. Superior Tribunal de Justiça. *Comunicado da Presidência do STJ*. Brasília, 19 Nov. 2020. Available at: https://www.stj.jus.br/sites/portalp/Paginas/Comunicacao/Noticias/19112020-Comunicado-da-Presidência-do-STJ.aspx. Accessed on: Oct. 24, 2025.

CÂMARA JÚNIOR, A. T. da. *Processamento de linguagem natural para indexação automática semântico-ontológica*. 2013. Doctoral Thesis (Doctorate in Information Science) – Universidade de Brasília, Brasília, 2013.

DATATILSYNET. *Artificial intelligence and privacy*. [S. l.], 2018.

FILIPOVA, I. A. Legal regulation of artificial intelligence: Experience of China. *Journal of Digital Technologies and Law*, [S. l.], v. 2, n. 1, p. 46-73, 2024.

FILTZ, E. et al. Interlinking Legal Data. *In*: INTERNATIONAL CONFERENCE ON SEMANTIC SYSTEMS, 14th., 2018, [S. l.]. *Proceedings...* [S. l.]: CEUR-WS.org, 2018. p. 1-4. Available at: https://www.researchgate.net/publication/344378847_Interlinking_Legal_Data. Accessed on: Oct. 24, 2025.

FUNDAÇÃO GETULIO VARGAS. *Inteligência artificial: tecnologia aplicada à gestão dos conflitos no âmbito do Poder Judiciário brasileiro*. Rio de Janeiro: FGV Justiça, Out. 2024. Available at: https://justica.fgv.br/sites/default/files/2024-10/ia_2fase_0.pdf. Accessed on: Oct. 24, 2025.

HOBBS, J. R.; RILOFF, E. Information extraction. *In*: INDURKHYA, N.; DAMERAU, F. J. (ed.). *Natural Language Processing*. 2nd ed. [S. l.]: Chapman & Hall/CRC, 2010. p. 511-532.

JENKINS, J. What can information technology do for law? *Harvard Journal of Law & Technology*, [S. l.], v. 21, n. 2, p. 589-607, 2008.

KATZ, D. M. et al. *Natural language processing in the legal domain*. [S. l.]: arXiv, 23 Feb. 2023. Preprint. Available at: https://arxiv.org/abs/2302.12039. Accessed on: Oct. 24, 2025.

LANCASTER, F. W. *Indexação e resumos: teoria e prática*. 2ª ed. [S. l.]: Briquet de Lemos, 2004.

MIGALHAS. *PJe ou e-Proc? Tribunais contestam resolução do CNJ sobre suspensão imediata de e-Proc*. Migalhas Quentes, 31 Out. 2019. Available at: https://www.migalhas.com.br/quentes/314284/pje-ou-e-proc--tribunais-contestam-resolucao-do-cnj-sobre-suspensao-imediata-de-e-proc. Accessed on: Oct. 24, 2025.

MORAIS, G. R. de. *Pressupostos e diretrizes para a aplicação do processamento de linguagem natural à recuperação de informações jurisprudenciais*: uma aplicação da AI4SG. [S. l.]: Editora Dialética, 2023.

MORAIS, G. R. de. Why natural language processing is indispensable to entrech Brazil's precedents system. *In*: GOMES, A. de O. et al. (ed.). *Encontro de Administração da Justiça*: Anais do ENAJUS 2021. [S. l.]: IBEPES, 2021.

NASTASE, V.; MIHALCEA, R.; RADEV, D. R. A survey of graphs in natural language processing. *Natural Language Engineering*, [S. l.], v. 21, n. 5, p. 665-698, 2015.

NUNES, D.; MARQUES, A. L. P. C. Algoritmo: o risco da decisão das máquinas. *Bonijuris*, [S. l.], v. 31, n. 659, p. 44-58, 2019.

PALMER, D. D. Text preprocessing. *In*: INDURKHYA, N.; DAMERAU, F. J. (ed.). *Natural Language Processing*. 2nd ed. [S. l.]: Chapman & Hall/CRC, 2010. p. 9-30.

PUENTE, M.; COALLA, L. *Mineração de Dados Textuais: Conceitos e Técnicas*. [S. l.]: Blucher, 2023.

SALTZER, J.; SCHROEDER, M. *The protection of information in computer systems*. [S. l.]: University of Virginia, 1975. Available at: https://www.cs.virginia.edu/~evans/cs551/saltzer/. Accessed on: Oct. 24, 2025.

SAVOY, J.; GAUSSIER, E. Information retrieval. *In*: INDURKHYA, N.; DAMERAU, F. J. (ed.). *Natural Language Processing*. 2nd ed. [S. l.]: Chapman & Hall/CRC, 2010. p. 455-484.

SILVA, N. C. da et al. Document type classification for Brazil's supreme court using a Convolutional Neural Network. *In*: THE TENTH INTERNATIONAL CONFERENCE ON FORENSIC COMPUTER SCIENCE AND CYBER LAW, 2018. *Proceedings...* [S. l.]: ICOFCS, 2018. p. 7-11.

TAN, A.-H. Text mining: The state of the art and the challenges. *In*: WORKSHOP ON KNOWLEDGE DISCOVERY FROM ADVANCED DATABASES, 1999. *Annals...* [S. l.]: [s. n.], 1999. p. 65-70.

WINKELS, R. Automatic classification of civil law cases. *In*: FERNANDES, R. V. C.; CARVALHO, A. G. P. de (ed.). *Tecnologia Jurídica & Direito Digital*: II Congresso Internacional de Direito, Governo e Tecnologia. [S. l.]: Fórum, 2018. p. 71-81.

YAROWSKY, D. Word sense disambiguation. *In*: INDURKHYA, N.; DAMERAU, F. J. (ed.). *Natural Language Processing*. 2nd ed. [S. l.]: Chapman & Hall/CRC, 2010. p. 315-338.

VI
ARTIFICIAL INTELLIGENCE IN REGULATED SECTORS AND THE DIGITAL ECONOMY

COPYRIGHT FOR THE INPUTS OF ARTIFICIAL INTELLIGENCE: BALANCING FREE ENTERPRISE WITH THE RIGHTS OF AUTHORS

Nicolo Zingales

Professor, FGV Direito Rio

Gustavo Salema Marques

PhD Candidate, Universidade Federal do Rio de Janeiro

DOI: https://doi.org/10.59224/dti8.ch20

Abstract: This article seeks to identify which comparative regulatory model is best suited to the Brazilian context for regulating the use of copyrighted works in the training of generative AI models. To this end, it starts from the conflict between copyright and free enterprise in order to advocate for the adoption of a flexible framework capable of allowing case-by-case analysis and avoiding the automatic predominance of either authors' or developers' interests. Finally, it proposes that the identified model be interpreted through the lens of fundamental rights and the social function of property.

Keywords: Regulation; Generative AI; Text and Data Mining.

Table of contents: 1. Introduction. 2. Generative AI training and the copyright problem. 2.1 expressive works and the generative AI supply chain. 3. Regulatory responses to text and data mining. 3.1. Global regulatory overview. 3.2. U.S. And South Korea's fair use and its application in Brazil. 3.2.1. What is fair use and how does it work. 3.2.2. Comparative overview of U.S. And South Korean fair use provisions. 3.2.3. Comparative overview of U.S. and South Korean fair use guidelines. 4. Fair use as a regulatory strategy for AI training in Brazil. 5. Conclusion. References

1. INTRODUCTION

Generative artificial intelligence (AI) is a category of artificial intelligence capable of generating new content, such as text, images, and music, based on text, audio, or image prompts, using as reference the dataset on which the model was trained.[1]. Accordingly, generative AI models rely on transformative technologies that require large volumes of data to achieve the objective of content generation, simulating human creativity in order to assist in text writing, artistic creation, and musical composition.[2]

Given their reliance on vast amounts of data and on extracting the value embedded within such data, the process of text and data mining (TDM) is essential for the training of generative AI models[3], since this process enables the extraction of information embedded in data and the identification of patterns within datasets, thereby leading to the creation of new knowledge[4] through the copying of content contained in such data.[5].

---

1 HACKER, Philipp; ENGEL, Andreas; HAMMER, Sarah; MITTELSTADT, Brent. *Introduction to the foundations and regulation of generative AI*. In: HACKER, Philipp; ENGEL, Andreas; HAMMER, Sarah; MITTELSTADT, Brent (org.). The Oxford handbook of the foundation and regulation of generative AI. Oxford: Oxford University Press, 2025. Available at: < https://ssrn.com/abstract=5137750 >. Accessed on: 26 May 2026. p. 2.

2 HACKER, Philipp; ENGEL, Andreas; HAMMER, Sarah; MITTELSTADT, Brent. *Introduction to the foundations and regulation of generative AI*. In: HACKER, Philipp; ENGEL, Andreas; HAMMER, Sarah; MITTELSTADT, Brent (org.). The Oxford handbook of the foundation and regulation of generative AI. Oxford: Oxford University Press, 2025. Available at: < https://ssrn.com/abstract=5137750 >. Accessed on: 26 May 2026. p. 2-3.

3 DERMAWAN, Artha. *Text and data mining exceptions in the development of generative AI models: what the EU member states could learn from the Japanese "nonenjoyment" purposes?*. The Journal of World Intellectual Property, v. 27, n. 1, p. 44-68, 2024. Available at: < https://onlinelibrary.wiley.com/doi/full/10.1111/jwip.12285 >. Accessed on: 26 May 2026. p. 49.

4 ZINGALES, Nicolo. *Of coffee pods, videogames, and missed interoperability: reflections for EU governance of the Internet of Things*. TILEC Discussion Paper, n. 2015-026, 1 dez. 2015. Available at: < https://ssrn.com/abstract=2707570 >. Accessed on: 26 May 2026. p. 4.

5 DERMAWAN, Artha. *Text and data mining exceptions in the development of generative AI models: what the EU member states could learn from the Japanese "nonenjoyment" purposes?*. The Journal of World Intellectual Property, v. 27, n. 1, p. 44-68, 2024. Available at: < https://onlineli-

However, during the training of generative AI models, there is extensive use of data related to works of human creativity protected by copyright, which are used by generative AI companies without authorization from copyright holders,[6] thereby raising the possibility of copyright infringement due to unauthorized reproductions occurring during the training process.[7]

Consequently, because the copies made during text and data mining have the potential to infringe copyright in the reproduced works, authors have expressed concern regarding the use of their works without authorization, arguing that generative AI companies not only violate copyright but also engage in unjust enrichment by failing to compensate the authors whose works are used.[8]

Against this backdrop, legislators around the world have proposed regulatory models aimed at governing the use of copyrighted works in the training of generative AI models[9]. Legislative approaches to addressing the use of such works in AI training vary across jurisdictions. While some countries establish specific rules to restrict TDM, particularly for commercial purposes, others adopt flexible legal frameworks capable of assessing the context in which a work was used in order to determine

brary.wiley.com/doi/full/10.1111/jwip.12285 >. Accessed on: 26 May 2026. p. 47.

6 LEE, Katherine; COOPER, A. Feder; GRIMMELMANN, James. *Talkin'bout AI generation: copyright and the generative-AI supply chain*. Available at: < https://james.grimmelmann.net/files/articles/talkin-bout-ai-generation >. Accessed on: 26 May 2026. p. 3.

7 MANTEGNA, Micaela. *ARTEficial: Creatividad, Inteligencia Artificial y Derecho de Autor*. Buenos Aires: CDYT-CETyS, 2023.p. 76.

8 UNITED STATES OF AMERICA. *United States District Court Southern District of New York, Ação Coletiva nº. 1:23-cv-8282-SHS*. Authors Guild et. al. contra Open AI e Micorsoft Corporation. Available at: < https://authorsguild.org/app/uploads/2023/12/Authors-Guild-OpenAI-Microsoft-Class-Action-Complaint-Dec-2023.pdf >. Accessed on: 26 May 2026, p. 1-4.

9 UNITED STATES OF AMERICA. Copyright Office. *Copyright and Artificial Intelligence, Part 3: Generative AI Training*. Washington, DC, maio 2025. Available at: < https://www.copyright.gov/ai/Copyright-and-Artificial-Intelligence-Part-3-Generative-AI-Training-Report-Pre-Publication-Version.pdf >. Accessed on: 26 May 2026, p. 1.

whether such use infringes copyright holders' rights[10].

In this regard, it should be noted that the Brazilian proposal permits the use of protected works during text and data mining for purposes related to research and the development of artificial intelligence systems. Nevertheless, such permission is precluded when access to the works is unlawful, when their use is intended for commercial purposes, or when the use is carried out in a manner that harms the economic interests of authors[11].

Accordingly, it may be concluded that the Brazilian model for regulating the use of works in AI training tends toward the protection of authors' rights, expressly recognizing the impossibility of commercial use of works without remuneration to authors, thereby effectively creating an absolute right to compensation whenever their works are commercially exploited within developers' economic activities.

---

[10] UNITED STATES OF AMERICA. Copyright Office. *Copyright and Artificial Intelligence, Part 3: Generative AI Training*. Washington, DC, maio 2025. Available at: < https://www.copyright.gov/ai/Copyright-and-Artificial-Intelligence-Part-3-Generative-AI-Training-Report-Pre-Publication-Version.pdf >. Accessed on: 26 May 2026, p. 76-79.

[11] BRAZIL. *PL nº 2.338/2023*. Available at: < https://www.camara.leg.br/proposicoesWeb/fichadetramitacao?idProposicao=2487262 >. Accessed on: 26 May 2026. [Article 63. The automated use of protected content in text and data mining processes for the purpose of research and development of AI systems by scientific, research, and educational organizations and institutions, museums, public archives, and libraries shall not constitute an infringement of copyright and related rights, provided that the following conditions are met: I – access to the content has been obtained lawfully; II – the use is not for commercial purposes; III – the use of content protected by copyright and related rights is limited to the extent necessary to achieve the intended objective, without prejudice to the economic interests of rights holders and without competing with the normal exploitation of the protected works and content. § 1. Copies of content protected by copyright and related rights used in AI systems shall be stored under secure conditions and only for the period necessary to carry out the activity or for the specific purpose of verifying results. § 2. The display or dissemination of works and content protected by copyright and related rights used in the development of AI systems is prohibited. § 3. This Article shall not apply to institutions affiliated with, associated with, or controlled by for-profit entities that provide AI systems or that hold equity participation among themselves. § 4. The provisions of the *caput* of this Article shall also apply to data mining carried out by public or private entities in the context of AI systems designed to combat civil and criminal offenses that infringe copyright and related rights.]

However, copyright, as a fundamental right[12], should not be treated as an absolute and non-relativizable right[13]. Therefore, when it comes into conflict with the fundamental right to free enterprise[14] held by generative AI developers, copyright must be balanced so that, in the specific case, it may be determined which fundamental right should prevail.

In this sense, the purpose of this article is to identify, among the regulatory models proposed by other countries, a framework capable of enabling, within the Brazilian legal context, the creation of a legal provision that allows for balancing authors' rights with developers' freedom of enterprise.

Accordingly, the objective is to identify a regulatory strategy capable of guiding judicial decision-making in a manner that allows for an analysis based on the circumstances of each case, without automatically guaranteeing either authors' or developers' interests by law. This would prevent both developers' economic activities and authors' rights from being unjustly restricted in the absence of judicial determination concerning conflicts arising from the use of copyrighted works in generative AI training.

Thus, through an analysis of foreign legal frameworks and considering the characteristics of generative AI training, this article seeks to answer the following question: which comparative model is best suited to the Brazilian context for enabling case-by-case judicial analysis?

To this end, the article first examines how generative AI models are trained and how such training may lead to copyright infringement. It then presents a regulatory overview to understand how different countries have sought to regulate the use of copyrighted works during generative AI training. Subsequently, a comparative anal-

---

[12] BRAZIL. *Constituição de República Federativa do Brasil de 1988*. Available at: < https://www.planalto.gov.br/ccivil_03/constituicao/constituicaocompilado.htm> Accessed on: 26 May 2026. Art. 5º, XXVII.

[13] BULOS, Uadi Lamêgo. *Curso de Direito Constitucional*. 11 ed. São Paulo: Saraiva Educação, 2018, p. 534.

[14] BRAZIL. *Constituição de República Federativa do Brasil de 1988*. Available at: < https://www.planalto.gov.br/ccivil_03/constituicao/constituicaocompilado.htm> Accessed on: 26 May 2026. Art. 5º, XIII e 170, parágrafo único.

ysis of the two regulatory models argued to be the most promising for Brazil is undertaken. Finally, the article analyzes how the proposed model may be adapted to the Brazilian legal context.

2. GENERATIVE AI TRAINING AND THE COPYRIGHT PROBLEM

The term artificial intelligence began to be used by a group of researchers who, in 1956, organized the workshop known as the "Dartmouth Summer Research Project on Artificial Intelligence[15]. From that moment onward, the development of artificial intelligence experienced a series of periods of enthusiasm and disappointment, since although the potential of AI could already be envisioned, only within the last ten years has the necessary context emerged for the development of this technology in its current form.[16]

Accordingly, it is important to note that the vast amount of data available on the internet, the increased capacity for data processing, the creation of new algorithms, and advances in machine learning techniques have been essential to the current stage of artificial intelligence development.[17]

With regard to machine learning, this technique consists of exposing an algorithm to a vast dataset with the aim of enabling it to perform a systematic and statistical analysis of patterns present within the dataset[18], thereby allowing it to acquire new capabilities, develop its own rules, and learn how to deal with novel situations.[19]

---

[15] CORTIZ, Diogo, *Inteligência artificial: conceitos fundamentais*, pp. 45-60 IN: VAINZOF, Rony e GUTIERREZ, Andrei (coord.). Inteligência Artificial: Sociedade, Economia e Estado. São Paulo: Revista dos Tribunais, 2021.p. 45-46.

[16] CORTIZ, Diogo, *Inteligência artificial: conceitos fundamentais*, pp. 45-60 IN: VAINZOF, Rony e GUTIERREZ, Andrei (coord.). Inteligência Artificial: Sociedade, Economia e Estado. São Paulo: Revista dos Tribunais, 2021.p. 45-46.

[17] CORTIZ, Diogo, *Inteligência artificial: conceitos fundamentais*, pp. 45-60 IN: VAINZOF, Rony e GUTIERREZ, Andrei (coord.). Inteligência Artificial: Sociedade, Economia e Estado. São Paulo: Revista dos Tribunais, 2021.p. 46-47.

[18] MANTEGNA, Micaela. *ARTEficial: Creatividad, Inteligencia Artificial y Derecho de Autor*. Buenos Aires: CDYT-CETyS, 2023.p. 43.

[19] CORTIZ, Diogo, *Inteligência artificial: conceitos fundamentais*, pp. 45-60 IN: VAINZOF, Rony e

In this context, concerning generative artificial intelligence (GAI), it may be argued that technological advances and the widespread availability of data have created a fertile environment for the development of this category of AI, given that both machine learning and internet-based data are essential to the construction of generative AI models[20].

Accordingly, it should be clarified that generative artificial intelligence models are trained with the purpose of generating creative works, approximating as closely as possible the creative capacity of human beings, whether through the production of texts, audio, or images.[21] However, for a GAI model to be effective in content generation, it is essential that it be trained through access to vast quantities of data[22], enabling it to capture the complexity and richness of information collected from the internet[23] and thereby become capable of imitating human creativity.

Therefore, considering that generative artificial intelligence models are trained through machine learning using billions of data points collected from the internet, which may correspond to works protected by copyright, questions arise regarding the possibility of copyright infringement occurring during the training of GAI mod-

GUTIERREZ, Andrei (coord.). Inteligência Artificial: Sociedade, Economia e Estado. São Paulo: Revista dos Tribunais, 2021.p. 45-46.

[20] LEE, Katherine, COOPER, A Feder e GRIMMELMANN, James. *Talkin' 'Bout AI Generation: Copyright and the Generative-AI Supply Chain.* Available at: < https://james.grimmelmann.net/files/articles/talkin-bout-ai-generation>. Accessed on: 26 May 2026. p. 14-15.

[21] LUCCHI, Nicola. *ChatGPT: A Case Study on Copyright Challenges for Generative Artificial Intelligence Systems*. Available at: < https://www.cambridge.org/core/journals/european-journal-of-risk-regulation/article/chatgpt-a-case-study-on-copyright-challenges-for-generative-artificial-intelligence-systems/CEDCE34DED599CC4EB201289BB161965 >. Accessed on: 26 May 2026.p. 1, 2 e 13.

[22] LUCCHI, Nicola. *ChatGPT: A Case Study on Copyright Challenges for Generative Artificial Intelligence Systems*. Available at: < https://www.cambridge.org/core/journals/european-journal-of-risk-regulation/article/chatgpt-a-case-study-on-copyright-challenges-for-generative-artificial-intelligence-systems/CEDCE34DED599CC4EB201289BB161965 >. Accessed on: 26 May 2026.p. 11.

[23] LEE, Katherine, COOPER, A Feder e GRIMMELMANN, James. *Talkin' 'Bout AI Generation: Copyright and the Generative-AI Supply Chain.* Available at: < https://james.grimmelmann.net/files/articles/talkin-bout-ai-generation>. Accessed on: 26 May 2026. p. 19.

els.

To address this issue, it is necessary to understand how generative artificial intelligence is developed, particularly during the training stage, which constitutes the central focus of this article, in order to clarify whether there is potential infringement of authors' rights during the learning process of generative AI models.

2.1 Expressive Works and the Generative AI Supply Chain

According to Katherine de Lee et al., Feder A. Coopere, and James Grimmelmann, the generative artificial intelligence supply chain does not begin with the training of a generative AI model, but rather with human creation itself[24]. Accordingly, the starting point for the development of this technology lies in the creation of expressive works, which may either originate in digital environments or be produced in physical form and subsequently digitized[25].

For the purposes of this article, it is important to understand that generative AI models are capable of generating content only because they have been trained on vast quantities of data derived from works of human creativity which, being available on the internet, are utilized throughout the generative AI supply chain[26].

In this regard, it should be noted that the generative artificial intelligence supply chain comprises eight stages: (i) The Creation of Expressive Works; (ii) Data Creation; (iii) Dataset Collection and Curation; (iv) Model (Pre-)Training; (v) Model Fine-Tuning; (vi) Model Release and System Deployment; (vii) Generation e (viii) Model Alignment[27].

---

[24] LEE, Katherine, COOPER, A Feder e GRIMMELMANN, James. *Talkin' 'Bout AI Generation: Copyright and the Generative-AI Supply Chain.* Available at: < https://james.grimmelmann.net/files/articles/talkin-bout-ai-generation>. Accessed on: 26 May 2026. p. 31.

[25] LEE, Katherine, COOPER, A Feder e GRIMMELMANN, James. *Talkin' 'Bout AI Generation: Copyright and the Generative-AI Supply Chain.* Available at: < https://james.grimmelmann.net/files/articles/talkin-bout-ai-generation>. Accessed on: 26 May 2026. p. 31-33.

[26] LEE, Katherine, COOPER, A Feder e GRIMMELMANN, James. *Talkin' 'Bout AI Generation: Copyright and the Generative-AI Supply Chain.* Available at: < https://james.grimmelmann.net/files/articles/talkin-bout-ai-generation>. Accessed on: 26 May 2026. p. 30-33.

[27] LEE, Katherine, COOPER, A Feder e GRIMMELMANN, James. *Talkin' 'Bout AI Generation:*

However, considering the scope of this chapter, the proposed analysis will be limited to the stages directly related to training itself. Accordingly, the first three stages are necessary to understand how human works are converted into data and subsequently organized into datasets, that is, how a phase of collection and preparation takes place, determining which data will be used in the training of a specific model[28].

Only after this process of data curation and organization does training begin. In this sense, both the training of a foundation model and its optimization essentially follow the same process: the developer selects the dataset most appropriate to the objectives of the model to be created, together with the training algorithm and the relevant training parameters. [29].

The training of AI models requires copying the content contained within data in order to identify patterns across datasets that may facilitate the discovery of new knowledge, a process made possible through the use of text and data mining (TDM) techniques[30].

Accordingly, irrespective of debates concerning copyright infringement, it is a fact that the need to analyze massive quantities of data makes the use of tools capable of processing such information indispensable, since the value of data lies not in the data itself but rather in the information extracted from it, thereby rendering the use of TDM imperative within the generative AI supply chain.[31]

Copyright and the Generative-AI Supply Chain. Available at: < https://james.grimmelmann.net/files/articles/talkin-bout-ai-generation>. Accessed on: 26 May 2026. p. 286.

[28] LEE, Katherine, COOPER, A Feder e GRIMMELMANN, James. *Talkin' 'Bout AI Generation: Copyright and the Generative-AI Supply Chain.* Available at: < https://james.grimmelmann.net/files/articles/talkin-bout-ai-generation>. Accessed on: 26 May 2026. p. 31-33.

[29] LEE, Katherine, COOPER, A Feder e GRIMMELMANN, James. *Talkin' 'Bout AI Generation: Copyright and the Generative-AI Supply Chain.* Available at: < https://james.grimmelmann.net/files/articles/talkin-bout-ai-generation>. Accessed on: 26 May 2026. p. 37-39.

[30] DERMAWAN, Artha. *Text and data mining exceptions in the development of generative AI models: What the EU member states could learn from the Japanese "nonenjoyment" purposes?*. The journal of World Intellectual Property. 2024, pp. 44-68. Available at: < https://onlinelibrary.wiley.com/doi/full/10.1111/jwip.12285> Accessed on: 26 May 2026. p. 47, 49 e 50.

[31] DERMAWAN, Artha. *Text and data mining exceptions in the development of generative AI models: What the EU member states could learn from the Japanese "nonenjoyment" purposes?*. The

Nevertheless, although text and data mining is essential for accessing information embedded in data, identifying patterns, and generating new knowledge [32], there remains the possibility of copyright infringement where the copies made during the TDM process reproduce copyrighted works without proper authorization [33].

In this regard, it must be recognized that the use of copyrighted works during the training of generative artificial intelligence models is undeniable, as such use is necessary for the development of models capable of generating content similar to that produced by humans [34]. Consequently, one may conclude that there is potential infringement of the author's exclusive right to authorize reproduction of their work whenever such use occurs outside the scope of legal exceptions.

Within the current Brazilian legal context, the use of a work without the author's authorization during the TDM process would only be permissible if it fell within one of the exceptions provided for in Articles 46–48 of Law No. 9,610 of February 19, 1998. These provisions include, for example, the reproduction of small excerpts of a work for private use and the partial reproduction of a work where such copying does not unjustifiably prejudice the rights of the author of the original work [35]. Therefore,

journal of World Intellectual Property. 2024, pp. 44-68. Available at: < https://onlinelibrary.wiley.com/doi/full/10.1111/jwip.12285> Accessed on: 26 May 2026. p. 49.

32 ZINGALES, Nicolo. *Of Coffee Pods, Videogames, and Missed Interoperability: Reflections for EU Governance of the Internet of Things* (December 1, 2015). TILEC Discussion Paper No. 2015-026, Available at: < https://ssrn.com/abstract=2707570 >. Accessed on: 26 May 2026.p. 4.

33 MANTEGNA, Micaela. *ARTEficial: Creatividad, Inteligencia Artificial y Derecho de Autor*. Buenos Aires: CDYT-CETyS, 2023.p. 76.

34 LUCCHI, Nicola. *ChatGPT: A Case Study on Copyright Challenges for Generative Artificial Intelligence Systems*. Disponível em: < https://www.cambridge.org/core/journals/european-journal-of-risk-regulation/article/chatgpt-a-case-study-on-copyright-challenges-for-generative-artificial-intelligence-systems/CEDCE34DED599CC4EB201289BB161965 >. Accessed on: 26 May 2026.p. 1, 2 e 13.

35 BRAZIL. *Lei 9.610 de fevereiro de 1998*. Available at: < https://www.planalto.gov.br/ccivil_03/leis/l9610.htm> Accessed on: 26 May 2026. [Art. 46. The following shall not constitute infringement of copyright: I – reproduction: a) in daily or periodical press, of news reports or informative articles published in newspapers or periodicals, provided that the author's name, if signed, and the source publication are indicated; b) in newspapers or periodicals, of speeches delivered at public meetings of any nature; c) of portraits or other forms of image representation made on commission, when reproduced by the owner of the commissioned object, provided there

if the copying carried out during text and data mining is not covered by these exceptions, it would be unlawful.

However, this remains a matter of ongoing debate. Although countries such as the United Kingdom recognize the existence of potential copyright infringement in this context [36], some scholars argue that the copying performed during the TDM process does not technically constitute reproduction for purposes of assessing copyright infringement, since what is reproduced is not the work itself but rather the

is no objection from the person portrayed or their heirs; d) of literary, artistic, or scientific works for the exclusive use of visually impaired persons, provided that such reproduction, without commercial purpose, is made through the Braille system or any other procedure and medium intended for these users; II – the reproduction, in a single copy of small excerpts, for the private use of the person making the copy, provided that it is carried out by such person and without profit motive; III – the quotation, in books, newspapers, magazines, or any other means of communication, of passages from any work for purposes of study, criticism, or debate, to the extent justified by the intended purpose, provided that the author's name and the source of the work are indicated; IV – the taking of lecture notes in educational institutions by those to whom the lessons are addressed, publication thereof, in whole or in part, being prohibited without the prior and express authorization of the lecturer; V – the use of literary, artistic, or scientific works, phonograms, and radio or television broadcasts in commercial establishments exclusively for demonstration to customers, provided that such establishments commercialize the media or equipment enabling their use; VI – theatrical performances and musical performances carried out within the family setting or, exclusively for educational purposes, in educational institutions, provided that in no case is there a profit motive; VII – the use of literary, artistic, or scientific works for the purpose of producing judicial or administrative evidence; VIII – the reproduction, in any work, of small excerpts from preexisting works of any nature, or of an entire work where involving visual arts, provided that reproduction itself is not the principal objective of the new work and does not prejudice the normal exploitation of the reproduced work nor cause unjustified harm to the legitimate interests of the authors. Art. 47. Paraphrases and parodies shall be freely permitted, provided that they do not constitute true reproductions of the original work nor bring it into disrepute. Art. 48. Works permanently situated in public places may be freely represented through paintings, drawings, photographs, and audiovisual means.]

[36] UNITED KINGDON (Intellectual Property Office). *Artificial Intelligence and Intellectual Property: copyright and patents: Government response to consultation*. (2022). Available at: < https://www.gov.uk/government/consultations/artificial-intelligence-and-ip-copyright-and-patents/outcome/artificial-intelligence-and-intellectual-property-copyright-and-patents-government-response-to-consultation#executive-summary>. Accessed on: 26 May 2026.

information contained within the data [37]. Accordingly, this latter position advocates for an update to copyright law aimed at defining what constitutes reproduction in the context of training generative artificial intelligence models, thereby providing greater legal certainty for both authors and developers.

As a result, scholarly debate has increasingly addressed the need to adapt copyright law to the emergence of generative artificial intelligence, with some authors even arguing that copyright may not be the most appropriate legal framework for regulating generative AI, given that other affected rights extend beyond the creative industries [38].

For purposes of delimiting the subject matter, however, this article will not engage extensively with that broader debate. Nevertheless, it raises a point central to the present study: the need to update copyright law in light of the development and advancement of generative AI.

Accordingly, this article does not seek to question the effectiveness of copyright law as a regulatory instrument. Rather, it argues for an evolution in its practical application so that authors' rights do not automatically prevail over the freedom of enterprise of generative AI developers.

In other words, the objective is to evaluate the most appropriate comparative model in order to propose the creation of a legal framework enabling the Brazilian judiciary to balance, on the basis of minimally objective standards, which fundamental right should prevail in a given case.

Accordingly, the following section presents a regulatory overview aimed at understanding which models have been adopted by jurisdictions seeking to regulate the use of text and data mining in the context of generative AI training, with the purpose of protecting the economic rights of authors and the creative industry as a whole.

---

[37] MANTEGNA, Micaela. *ARTificial: why copyright is not the right policy tool to deal with generative AI*. The Yale Law Journal Forum, 22 abr. 2024. Available at: < https://ssrn.com/abstract=5090127 >. Accessed on: 26 May 2026, p. 1160 e 1163.

[38] CRAIG, Carys. *The AI-Copyright Trap*. Osgoode Legal Studies Research Paper, n. 4905118, 15 jul. 2024. Available at: < https://ssrn.com/abstract=4905118 >. Accessed on: 26 May 2026, p.5.

3. REGULATORY RESPONSES TO TEXT AND DATA MINING

For the purpose of delimiting the scope of this study, this section does not seek to address all existing regulations concerning generative AI or even all legal frameworks governing text and data mining (TDM). Rather, its objective is to understand the regulatory paths that different jurisdictions are adopting in order to govern the use of protected works in the context of generative artificial intelligence training.

Accordingly, this regulatory overview is intended to make it possible to identify a model that promotes case-by-case analysis, as opposed to regulatory approaches that automatically privilege either authors' rights or developers' interests. Such identification will then provide the basis for a comparative analysis between this model, the existing legal framework, and the proposals currently under discussion in Brazil.

3.1. Global Regulatory Overview

For the purpose of delimiting this study, jurisdictions with existing or proposed regulations specifically aimed at governing generative AI were identified, and those directly addressing the use of text and data mining (TDM) for training purposes were selected. Accordingly, this section examines the legal frameworks of the European Union, the United Kingdom, Brazil, Japan, Singapore, Ecuador, South Korea, and the United States, all of which provide rules concerning the use of TDM in AI development.

At the outset, it may be observed that the central issue concerns the use of text and data mining for commercial purposes, revealing an economic dispute between copyright holders and the artificial intelligence sector, as well as the attempts made by jurisdictions either to balance these competing interests or to favor one side over the other..

This dynamic becomes particularly evident when examining the regulatory frameworks of the European Union and the United Kingdom, as well as the Brazilian legislative proposal. Regarding the European Union, it was found that the relevant institutions chose to regulate artificial intelligence, including generative models, through a risk-based approach, focusing on protecting economic sectors potentially

affected by AI[39].

Accordingly, concern for copyright protection can already be observed in the European Union's artificial intelligence framework (AI Act), which simultaneously recognizes the importance of generative AI and the promotion of innovation. Nevertheless, there is a clear concern with preserving the interests of authors and artists in light of the potential unauthorized use of their works during text and data mining processes[40].

Although this concern with innovation exists, when conflicts arise with the economic rights of authors, these rights ultimately receive preferential treatment under Directive (EU) 2019/790, which specifically addresses the use of works in text and data mining processes.[41].

---

[39] MARQUES, Gustavo Salema. *Estratégias para flexibilização do regime de proteção dos direitos autorais: proposta para viabilizar a mineração de textos e dados (TDM) no treinamento de modelos de inteligência artificial generativa*. Dissertation (Master's Degree in Regulatory Law) — Fundação Getulio Vargas, Escola de Direito do Rio de Janeiro, Rio de Janeiro, 2025, p. 77.

[40] EUROPEAN UNION. *Regulation (EU) 2024/1689 of the European Parliament and of The Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act).* Available at: < https://eur-lex.europa.eu/eli/reg/2024/1689/oj >. Accessed on: 26 May 2026. [(105). General-purpose AI models, in particular large generative AI models, capable of generating text, images, and other content, present unique innovation opportunities but also challenges to artists, authors, and other creators and the way their creative content is created, distributed, used and consumed. The development and training of such models require access to vast amounts of text, images, videos and other data. Text and data mining techniques may be used extensively in this context for the retrieval and analysis of such content, which may be protected by copyright and related rights. Any use of copyright protected content requires the authorisation of the rightsholder concerned unless relevant copyright exceptions and limitations apply. Directive (EU) 2019/790 introduced exceptions and limitations allowing reproductions and extractions of works or other subject matter, for the purpose of text and data mining, under certain conditions. Under these rules, rightsholders may choose to reserve their rights over their works or other subject matter to prevent text and data mining, unless this is done for the purposes of scientific research. Where the rights to opt out has been expressly reserved in an appropriate manner, providers of general-purpose AI models need to obtain an authorisation from rightsholders if they want to carry out text and data mining over such works.]

[41] ZIAJA, Gina Maria. *The text and data mining opt-out in Article 4(3) CDSMD: Adequate veto*

It is important to highlight that Article 4(1) of Directive (EU) 2019/790 establishes an exception to copyright capable of permitting the use of protected works in text and data mining procedures irrespective of the nature of the use, thereby allowing such activities even for commercial purposes.[42].

However, Article 4(3) introduces an opt-out mechanism for copyright holders, allowing them to prohibit the use of their works for training artificial intelligence models[43]. This mechanism ultimately renders the development of generative AI models excessively costly, since developers may be required to compensate each author whose work is used, and, in a context involving billions of data points, such costs may become prohibitive.[44].

By creating an opt-out regime, the European Union affects not only the economic viability of the generative AI sector but also its own competitiveness within the global market[45]. Accordingly, depending on its practical application, the EU may place itself

right for rightholders or a suffocating blanket for European artificial intelligence innovations? Available at: < https://academic.oup.com/jiplp/article/19/5/453/7614898 >. Accessed on: 26 May 2026.p. 455.

42 EUROPEAN UNION. *Directive (EU) 2019/790 of the European Parliament and of the Council of 17 April 2019 on copyright and related rights in the Digital Single Market* Available at: < https://eur-lex.europa.eu/eli/dir/2019/790/oj >. Accessed on: 26 May 2026. [Article 4 (1) Member States shall provide for an exception or limitation to the rights provided for in Article 5(a) and Article 7(1) of Directive 96/9/EC, Article 2 of Directive 2001/29/EC, Article 4(1)(a) and (b) of Directive 2009/24/EC and Article 15(1) of this Directive for reproductions and extractions of lawfully accessible works and other subject matter for the purposes of text and data mining.]

43 EUROPEAN UNION. *Directive (EU) 2019/790 of the European Parliament and of the Council of 17 April 2019 on copyright and related rights in the Digital Single Market* Available at: < https://eur-lex.europa.eu/eli/dir/2019/790/oj >. Accessed on: 26 May 2026. [Article 4 (3) The exception or limitation provided for in paragraph 1 shall apply on condition that the use of works and other subject matter referred to in that paragraph has not been expressly reserved by their rightholders in an appropriate manner, such as machine-readable means in the case of content made publicly available online.]

44 ZIAJA, Gina Maria. *The text and data mining opt-out in Article 4(3) CDSMD: Adequate veto right for rightholders or a suffocating blanket for European artificial intelligence innovations?* Available at: < https://academic.oup.com/jiplp/article/19/5/453/7614898 >. Accessed on: 26 May 2026.p. 455.

45 ZIAJA, Gina Maria. *The text and data mining opt-out in Article 4(3) CDSMD: Adequate veto*

at a competitive disadvantage relative to other jurisdictions by establishing a framework capable of restricting innovation and AI development.[46].

A similar scenario may emerge in Brazil, given that Brazilian lawmakers relied upon the European model of AI regulation[47] and adopted regulatory strategies such as the opt-out mechanism embodied in Article 64 of Legislative Proposal No. 2,338/2023[48]. Like its European counterpart, this provision has the potential to hinder developers' freedom of enterprise, placing Brazil at a disadvantage in relation to other countries and thereby limiting its capacity for economic growth through innovation in the generative AI sector.

In this regard, it should be noted that Brazil, unlike the European Union, has chosen to prohibit the use of TDM for commercial purposes absent remuneration of copyright holders whose works are used in AI training. This decision, in itself, may negatively affect the country's economic growth, given that the possibility of commercial TDM has the potential to generate new products and services capable of stimulating national economies[49].

Compounding this issue, Brazilian legislators, recognizing the importance of the entertainment sector to the country, have sought to provide greater protection to this

---

right for rightholders or a suffocating blanket for European artificial intelligence innovations? Available at: < https://academic.oup.com/jiplp/article/19/5/453/7614898 >. Accessed on: 26 May 2026.p. 455.

46 ZIAJA, Gina Maria. *The text and data mining opt-out in Article 4(3) CDSMD: Adequate veto right for rightholders or a suffocating blanket for European artificial intelligence innovations?* Available at: < https://academic.oup.com/jiplp/article/19/5/453/7614898 >. Accessed on: 26 May 2026.p. 455 e 459.

47 SOUZA, Allan Rocha de (coord.) *Inteligência Artificial e Direitos Autorais: Contribuições ao debate regulatório no Brasil*. Rio de Janeiro: IBDAutoral, 2024. Available at: < https://ibdautoral.org.br/resources/IA-E-DIREITOS-AUTORAIS_v2.pdf >. Accessed on: 26 May 2026. p. 90.

48 BRAZIL. *PL nº 2.338/2023*. Available at: < https://www.camara.leg.br/proposicoesWeb/fichadetramitacao?idProposicao=2487262 >. Accessed on: 26 May 2026. [art. Art. 64. O titular de direitos de autor e conexos poderá proibir a utilização dos conteúdos de sua titularidade no desenvolvimento de sistemas de IA nas hipóteses não contempladas pelo art. 63 desta Lei.]

49 ZINGALES, Nicolo. *Of Coffee Pods, Videogames, and Missed Interoperability: Reflections for EU Governance of the Internet of Things* (December 1, 2015). TILEC Discussion Paper No. 2015-026, Available at: < https://ssrn.com/abstract=2707570 >. Accessed on: 26 May 2026.p. 4.

segment of the economy in order to mitigate the risks posed by generative AI to the creative industries, viewing the protection of cultural industries as a commitment linked to both the economy and national identity[50].

Such a position, however, has led lawmakers to adopt norms that substantially burden developers' activities, such as Articles 62 and 65 of Bill No. 2,338/2019, which respectively establish auditing requirements and a burdensome compensation regime that may prove impracticable for developers to satisfy[51].

Accordingly, the Brazilian legislature's preference for protecting creative economy organizations over generative AI developers becomes evident, demonstrating a deliberate effort to safeguard the economic interests of copyright-holding entities even where this results in disproportionate and onerous regulation capable of restricting the freedom of enterprise of those engaged in developing generative AI models.

Such protectionism may be understandable given that the leading generative AI models are currently headquartered outside Brazil. Nevertheless, the argument advanced here is that the legitimate need to protect Brazil's cultural sector should not prevent the creation of a domestic generative AI industry capable of making the country competitive in the global market.

It should also be noted that the Brazilian scenario is not unique. Other jurisdictions likewise face significant pressure from the creative industries and consider their protection justified. The United Kingdom, for example, proposed creating an exception permitting the use of protected works during TDM processes precisely to establish itself as a global leader in AI development[52].

---

50 BRAZIL. *Emenda nº 191 à Proposta legislativa nº 2.338 de 2023*. Senado Federal. Senador Efraim Filho. Available at: < https://legis.senado.leg.br/sdleg-getter/documento?dm=9854105&ts=1734002258479&disposition=inline&ts=1734002258479 > Accessed on: 26 May 2026.

51 MARQUES, Gustavo Salema. *Estratégias para flexibilização do regime de proteção dos direitos autorais: proposta para viabilizar a mineração de textos e dados (TDM) no treinamento de modelos de inteligência artificial generativa*. Dissertation (Master's Degree in Regulatory Law) — Fundação Getulio Vargas, Escola de Direito do Rio de Janeiro, Rio de Janeiro, 2025, p. 89 e 93.

52 MARQUES, Gustavo Salema. *Estratégias para flexibilização do regime de proteção dos direitos autorais: proposta para viabilizar a mineração de textos e dados (TDM) no treinamento de modelos*

However, because the proposal would have allowed commercial text and data mining without remuneration to copyright holders, the UK Parliament, following pressure from entertainment companies, persuaded the Executive to abandon the initiative[53]. This clearly reflects lawmakers' preference for protecting the economic interests of the creative industries, despite the AI sector's argument that requiring prior authorization for the use of works in TDM processes would render developers' activities excessively burdensome.[54].

Legislators in Ecuador and Singapore have followed a similar path. Although both jurisdictions possess flexible legal rules capable of enabling case-by-case analysis, they nevertheless chose to prohibit the use of creative works in text and data mining for commercial purposes [55].

Accordingly, these two countries adopted dual frameworks: a general and flexible rule modeled after American fair use, combined with a specific prohibition on the unauthorized use of works where TDM is employed for commercial purposes, thereby creating an automatic barrier to AI development without allowing judicial balancing [56].

---

de inteligência artificial generativa. Dissertation (Master's Degree in Regulatory Law) — Fundação Getulio Vargas, Escola de Direito do Rio de Janeiro, Rio de Janeiro, 2025, p. 55.

53 MARQUES, Gustavo Salema. *Estratégias para flexibilização do regime de proteção dos direitos autorais: proposta para viabilizar a mineração de textos e dados (TDM) no treinamento de modelos de inteligência artificial generativa*. Dissertation (Master's Degree in Regulatory Law) — Fundação Getulio Vargas, Escola de Direito do Rio de Janeiro, Rio de Janeiro, 2025, p. 52-58 e 78.

54 MARQUES, Gustavo Salema. *Estratégias para flexibilização do regime de proteção dos direitos autorais: proposta para viabilizar a mineração de textos e dados (TDM) no treinamento de modelos de inteligência artificial generativa*. Dissertation (Master's Degree in Regulatory Law) — Fundação Getulio Vargas, Escola de Direito do Rio de Janeiro, Rio de Janeiro, 2025, p. 52-58 e 78.

55 MARQUES, Gustavo Salema. *Estratégias para flexibilização do regime de proteção dos direitos autorais: proposta para viabilizar a mineração de textos e dados (TDM) no treinamento de modelos de inteligência artificial generativa*. Dissertation (Master's Degree in Regulatory Law) — Fundação Getulio Vargas, Escola de Direito do Rio de Janeiro, Rio de Janeiro, 2025, p. 73-77 e 80.

56 MARQUES, Gustavo Salema. *Estratégias para flexibilização do regime de proteção dos direitos autorais: proposta para viabilizar a mineração de textos e dados (TDM) no treinamento de modelos de inteligência artificial generativa*. Dissertation (Master's Degree in Regulatory Law) — Fundação Getulio Vargas, Escola de Direito do Rio de Janeiro, Rio de Janeiro, 2025, p. 73-77 e 80.

In light of this, it is important to emphasize that this study does not endorse the notion that legal norms should automatically grant developers unrestricted rights to use creative works, nor does it advocate principles that unjustifiably hinder the protection of authors' rights.

Regarding the Japanese solution, the principle of non-enjoyment initially appears attractive. Under this principle, if the use of a work does not involve enjoyment of the thoughts or emotions expressed therein, there would be no copyright infringement unless such use unjustifiably harms the author's rights.[57].

The Japanese government has characterized this principle as a copyright exception specifically designed to facilitate the use of works in text and data mining processes, including for commercial purposes.[58].

However, because this exception is particularly broad, authors' rights receive protection only where unjustified harm can be demonstrated. Since machine reading does not, in any meaningful sense, enjoy or appreciate the thoughts or emotions expressed in a work, questions arise as to whether the principle of non-enjoyment

---

[57] JAPAN. *Copyright Law of Japan*. Available at: < https://www.cric.or.jp/english/clj/cl1.html.>. Accessed on: 26 May 2026. [Artigo 30-4 A work may be exploited in any manner and to the extent necessary in the following cases, or in other cases where the purpose is neither to personally enjoy nor to cause others to enjoy the thoughts or sentiments expressed in the work; however, this shall not apply where such exploitation unjustifiably prejudices the interests of the copyright holder, taking into account the nature or purpose of the work and the circumstances surrounding its exploitation: (i) for testing related to the development or practical application of technologies associated with the recording of sounds or images of a work, or with other similar forms of exploitation; (ii) for data analysis (understood as the extraction, comparison, classification, or other statistical analysis of linguistic, sound, visual, or other elements derived from a large number of works or comparable data); (iii) for computer data processing or other forms of exploitation that do not involve the perception of the expressive elements of the work through human senses (in the case of computer programs, this excludes execution of the program on a computer), in addition to the uses described in the preceding items.]

[58] JAPAN. Agency for Cultural Affairs, Government of Japan. *General Understanding on AI and Copyright in Japan*. Disponível em: < https://www.bunka.go.jp/english/policy/copyright/pdf/94055801_01.pdf >. Accessed on: 26 May 2026. p. 4.

may effectively result in automatic protection of developers' interests.[59].

Accordingly, it is legitimate to question whether adopting the Japanese rule without adaptation would reduce authors' ability to challenge the use of their works in generative AI training. Although broad, the rule does not initially appear sufficiently flexible to permit genuine case-by-case analysis, since it would never be possible to claim that an algorithm appreciated a work in the same manner as a human being.[60].

According to the South Korean government, although the creation of a TDM exception aimed at promoting the AI sector appeared attractive, the experience of countries adopting such approaches was ultimately viewed negatively [61]. The relevant commission concluded that exception-based rules in other jurisdictions had achieved limited success in reducing conflicts between authors and developers, leading South Korea to consider it preferable to formulate a solution grounded in existing fair use principles and through optimization of the current regulatory framework [62].

In this regard, fair use constitutes a general limitation on authors' economic rights that, within judicial proceedings and following the balancing of four statutory factors, permits third parties to use protected works where such use benefits the public interest, thereby dispensing with the need for prior authorization or remuneration [63].

---

59 SAG, Matthew. *The New Legal Landscape for Text Mining and Machine Learning*. Available at: < https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3331606 >. Accessed on: 26 May 2026. p.6.

60 SAG, Matthew. *The New Legal Landscape for Text Mining and Machine Learning*. Available at: < https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3331606 >. Accessed on: 26 May 2026. p.6.

61 SOUTH KOREA. Ministry of Culture, Sports and Tourism. *Imminent release of the fair use guidelines on training generative AI with copyrighted works*. Seul, 27 nov. 2025. Available at: < https://www.mcst.go.kr/site/s_notice/press/pressView.jsp?pSeq=22130 >. Accessed on: 26 May 2026. p. 1.

62 SOUTH KOREA. Ministry of Culture, Sports and Tourism. *Imminent release of the fair use guidelines on training generative AI with copyrighted works*. Seul, 27 nov. 2025. Available at: < https://www.mcst.go.kr/site/s_notice/press/pressView.jsp?pSeq=22130 >. Accessed on: 26 May 2026. p. 1.

63 LIN, Patrick K., *Fair's Fair: How Public Benefit Considerations in the Fair Use Doctrine Can Patch Bias in Artificial Intelligence Systems*. 11 Indiana Journal of Law & Social Equality 229 (2023). Available at: < https://ssrn.com/abstract=4116483>. Accessed on: 26 May 2026.p. 235.

Fair use is therefore a flexible and open-ended rule capable of applying to any use, provided that the legal requirements are satisfied[64]. These requirements enable a holistic and evaluative interpretation of the statutory factors in order to determine whether the use made by a third party was fair and sufficient to limit copyright protection in the specific case[65].

Precisely because it enables case-by-case assessment and possesses the flexibility necessary to keep pace with technological advancement, this article suggests the adoption of a fair use framework to mediate conflicts between authors' rights and developers' interests[66] in determining whether commercially oriented text and data mining should fall within a fair use exception.

Accordingly, because the objective is to identify the most appropriate model for Brazil, and because global experience with specific statutory exceptions has proven unsatisfactory, this article argues that adopting similar exception-based rules would not represent the most suitable path for the Brazilian legal system. Instead, it proposes consideration of a fair use rule capable of promoting case-by-case analysis as an appropriate framework for resolving conflicts between authors and developers in Brazil.

Accordingly, the following section examines the legal provisions found in South Korean and American law with the aim of determining which would provide the most appropriate model for the creation of a fair use rule capable of introducing greater flexibility into the Brazilian copyright system.

3.2. U.S. and South Korea's Fair Use and its application in Brazil

Based on the research conducted, specific guidelines were identified in the United

---

64 SOUZA, Allan Rocha de (coord.) *Inteligência Artificial e Direitos Autorais: Contribuições ao debate regulatório no Brasil*. Rio de Janeiro: IBDAutoral, 2024. Available at: < https://ibdautoral.org.br/resources/IA-E-DIREITOS-AUTORAIS_v2.pdf>. Accessed on: 26 May 2026. p. 26.

65 ASCENSÃO, José de Oliveira. *Estudos de José de Oliveira Ascensão sobre direito autoral e sociedade da informação*. Curitiba: IODA, 2022. p. 94-95.

66 LIN, Patrick K., *Fair's Fair: How Public Benefit Considerations in the Fair Use Doctrine Can Patch Bias in Artificial Intelligence Systems*. 11 Indiana Journal of Law & Social Equality 229 (2023). Available at: < https://ssrn.com/abstract=4116483>. Accessed on: 26 May 2026.p. 236.

States and South Korea aimed exclusively at guiding the analysis of fair use in the context of text and data mining for the training of artificial intelligence models. Accordingly, it should be clarified that no similar guidelines were identified in the other jurisdictions examined. For this reason, this section focuses solely on the legal provisions and interpretative guidance developed in the United States and South Korea concerning the application of fair use to the use of protected works in generative AI training.

Accordingly, this section is divided into three parts: (i) a brief explanation of the fair use doctrine and its practical application; (ii) a comparative analysis of the legal provisions governing fair use in the United States and South Korea; and (iii) a comparison between the guidance issued by the United States Copyright Office and the Korea Copyright Commission regarding judicial interpretation of fair use.

3.2.1. What is fair use and how does it works

Fair use is a judicial doctrine developed in the United States during the nineteenth century and codified in the Copyright Law of the United States in 1976, becoming a vital component of the American copyright protection system[67]. Since then, the doctrine has been shaped by American courts with the purpose of balancing the rights of authors against those of individuals who use protected works without prior authorization.[68]

Accordingly, in order to determine whether the use of a work may qualify as fair use, a judge must evaluate the four factors set forth in Section 107 of the Copyright Law of the United States. These factors must be considered and weighed against one another in order to determine, in the specific case, whether the use of the work was fair.[69]

---

67 BAINBRIDGE, Olivia., *Is the Fair Use Doctrine a viable argument for Generative AI?*. Available at: < https://scholarworks.wmich.edu/honors_theses/3756/ >. Accessed on: 26 May 2026.p. 7.

68 BAINBRIDGE, Olivia., *Is the Fair Use Doctrine a viable argument for Generative AI?*. Available at: < https://scholarworks.wmich.edu/honors_theses/3756/ >. Accessed on: 26 May 2026.p. 7.

69 BAINBRIDGE, Olivia., *Is the Fair Use Doctrine a viable argument for Generative AI?*. Available at: < https://scholarworks.wmich.edu/honors_theses/3756/ >. Accessed on: 26 May 2026.p. 8-9.

Fair use is therefore a flexible legal doctrine that requires both the assessment of the relevant factors and the outcome of such evaluation to be interpreted and balanced collectively, with the objective of determining whether an unauthorized use carried out by a third party is sufficiently justified to limit the author's exclusive rights over the work[70].

In this regard, it should be noted that the four fair use factors function as guiding principles for judicial decision-making. The judge must, in light of the circumstances of the specific case, determine whether an otherwise unauthorized use may be protected under fair use, with the burden resting upon the defendant to demonstrate that the use of the work falls within the boundaries of the doctrine.[71].

Having established this, with respect to the training of generative artificial intelligence, fair use constitutes a viable defense for developers where the relevant factors, following case-by-case judicial balancing, weigh in favor of their activities. Such an approach enables AI companies to defend their interests by demonstrating that their use of protected works may properly be characterized as fair use.[72]

3.2.2. Comparative Overview of U.S. and South Korean Fair Use Provisions

As previously explained, fair use is a doctrine originating in the United States that enables a contextualized assessment of the use of protected works, allowing developers to obtain legal authorization to use third-party works without prior permission.

In this regard, it should be noted that, in South Korea, fair use was incorporated into the South Korean Copyright Act in 2011 through legislative amendment, moti-

---

70 LIN, Patrick K., *Fair's Fair: How Public Benefit Considerations in the Fair Use Doctrine Can Patch Bias in Artificial Intelligence Systems*. 11 Indiana Journal of Law & Social Equality 229 (2023). Available at: < https://ssrn.com/abstract=4116483>. Accessed on: 26 May 2026.p. 236.

71 UNITED STATES OF AMERICA. Copyright Office. *Report on Copyright and Artificial Intelligence, Part 3: Generative AI Training*. Washington, DC, maio 2025. Available at: < https://www.copyright.gov/ai/Copyright-and-Artificial-Intelligence-Part-3-Generative-AI-Training-Report-Pre-Publication-Version.pdf >. Accessed on: 26 May 2026. p. 107.

72 BAINBRIDGE, Olivia., *Is the Fair Use Doctrine a viable argument for Generative AI?*. Available at: < https://scholarworks.wmich.edu/honors_theses/3756/ >. Accessed on: 26 May 2026.p. 29.

vated by the need to establish a legal provision capable of addressing situations in which specific limitation rules would not provide the level of effectiveness required by the digitization of works and the changing environments in which they circulate. [73]

Accordingly, the adoption of fair use within the South Korean copyright protection system was driven precisely by the need to respond to technological advances and to the evolving manner in which works circulate in digital environments, demonstrating that technological development requires a corresponding evolution in the way copyright law is applied.[74].

Against this background, a brief comparative analysis of the provisions contained in the Copyright Law of the United States and the South Korean Copyright Act is undertaken in order to understand the similarities and differences between the legal texts of both jurisdictions, as well as to assess whether the interpretation and application of South Korean fair use follows the same path as the American system.

Copyright Law of the United States	**South Korean Copyright Act**
§107 Limitations on exclusive rights: Fair use Notwithstanding the provisions of sections 106 and 106A, the fair use of a copyrighted work, including such use by reproduction in copies or phonorecords or by any other means specified by that section, for purposes such as criticism, comment, news reporting, teaching (including	Article 35-5 (Fair Use of Works) (1) Except as provided in Articles 23 through 35-4 and 101-3 through 101-5, where a person does not unreasonably undermine an author's legitimate interest without conflicting with the normal exploitation of works, he or she is entitled to use such works.

[73] SOUTH KOREA. *Guidelines on Fair Use under Copyright Law for the Training of Generative Artificial Intelligence on Copyrighted Works*. Available at:< https://www.copyright.or.kr/information-materials/publication/research-report/view.do?brdctsno=55211&pageIndex=1&noticeYn=&brdclasscodeList=&etc2=&etc1=&searchText=&searchkeyword=&brdclasscode=&nationcodeList=&searchTarget=ALL&nationcode=#>. Accessed on: 26 May 2026 p.26.

[74] MOUTINHO, Andréa Lasevicius; LIMA Isabela Pereira de Moura; FREIRE, Maria Eduarda Neuburguer; GUIMARÃES, Tatiana e SILVA, Alexandre Pacheco da. *Direito Autoral no Brasil e Internet: Revisando a História Recente à luz das Novas Tecnologias e plataformas digitais*. In: Direito Autoral e Internat: diagnóstico e perspectivas do debate brasileiro. Coordenação: MOUTINHO, Andréa Lasevicius; LIMA Isabela Pereira de Moura e SILVA, Alexandre Pacheco da. São Paulo: Editora Almedina, 2023.p 138.

multiple copies for classroom use), scholarship, or research, is not an infringement of copyright. In determining whether the use made of a work in any particular case is a fair use the factors to be considered shall include— (1) the purpose and character of the use, including whether such use is of a commercial nature or is for nonprofit educational purposes; (2) the nature of the copyrighted work; (3) the amount and substantiality of the portion used in relation to the copyrighted work as a whole; and (4) the effect of the use upon the potential market for or value of the copyrighted work. The fact that a work is unpublished shall not itself bar a finding of fair use if such finding is made upon consideration of all the above factors[75].	(2) In determining whether an act of using works falls under paragraph (1), the following matters shall be considered: 1. Purposes and characteristics of use; 2. Types and purposes of works; 3. Amount and substantiality of portion used in relation to the whole works; 4. Effect of the use of works on the existing or potential market for the works or current or potential value thereof[76].

Considering the provisions presented in the table above, it may be concluded that both systems pursue the same objective, namely, to permit the use of a protected work without causing harm to the market for the original work. Accordingly, although the provision contained in the Copyright Law of the United States is more detailed, both the consequences arising from the application of fair use and the factors to be considered are fundamentally the same.

With respect to the assessment of the four factors, South Korea follows the same logic as the United States, whereby the factors must be evaluated collectively rather than treated as independently determinative. Accordingly, if one factor weighs against fair use, it must nevertheless be interpreted in light of the overall balance of

---

75 UNITED STATES OF AMERICA. *Copyright Law of the United States (Title 17 United States Code.)*. Available at: < https://www.copyright.gov/title17/title17.pdf >. Accessed on: 26 May 2026. §107.

76 SOUTH KOREA. *Copyright Act. Act No. 17592, de 8 de dezembro de 2020*. Available at:< https://www.copyright.or.kr/eng/laws-and-treaties/copyright-law/act.do >. Accessed on: 26 May 2026. Article 35-5.

factors and should not be regarded as dispositive in and of itself.[77].

Accordingly, it may be concluded that South Korea, based on both the wording of its legal provisions and the interpretative guidance issued by the Korea Copyright Commission and the U.S. Copyright Office, adopts a normative and operational approach similar to that of the American system, establishing a general legal framework designed to address technological advancement.

3.2.3. Comparative Overview of U.S. and South Korean Fair Use Guidelines

Given that fair use has been increasingly regarded as a flexible general rule capable of addressing legal issues arising from technological development over recent decades, and considering that South Korea applies the doctrine in a manner similar to that of the United States, the analysis now turns to a comparative examination of how the U.S. Copyright Office and the Korea Copyright Commission understand the application of the four fair use factors in relation to the use of works within the context of generative artificial intelligence.

To this end, the following documents were examined: the Report on Copyright and Artificial Intelligence, Part 3: Generative AI Training issued by the U.S. Copyright Office and the Guidelines on Fair Use under Copyright Law for the Training of Generative Artificial Intelligence on Copyrighted Works issued by the Korea Copyright Commission. These materials provide guidance on how each fair use factor should be interpreted within their respective jurisdictions, as illustrated in the table below.

---

[77] SOUTH KOREA. *Guidelines on Fair Use under Copyright Law for the Training of Generative Artificial Intelligence on Copyrighted Works*. Available at::< https://www.copyright.or.kr/information-materials/publication/research-report/view.do?brdctsno=55211&pageIndex=1&noticeYn=&brdclasscodeList=&etc2=&etc1=&searchText=&searchkeyword=&brdclasscode=&nationcodeList=&searchTarget=ALL&nationcode=#>. Accessed on: 26 May 2026. p.26.

Factor	U. S. Copyright Office[78]	Korea Copyright Commission[79]
First factor (The purpose and character of the use)	**a) Transformative use:** Use will be considered transformative when training is conducted on a massive dataset that results in the creation of a model capable of generating outputs that neither substitute for nor reproduce the original work.. **b) Commercial nature:** Commercial use is a factor that weighs against a finding of fair use. However, the more transformative the use is, the less significance is attached to whether the use occurred for commercial or non-commercial purposes. **c) Legitimate access:**	**a) Transformative use:** The following elements favor a finding of fair use in this context: (i) training conducted in such a manner that the model generates outputs unrelated to the original works; (ii) the implementation of technological safeguards designed to prevent inputs capable of producing outputs substantially similar to the original works; and (iii) the development of a generative AI model that creates value and utility contributing to the public interest without adversely affecting the market for the original work. **b) Commercial nature:** The use of protected works for commercial purposes is regard-

78 UNITED STATES OF AMERICA. Copyright Office. *Report on Copyright and Artificial Intelligence, Part 3: Generative AI Training*. Washington, DC, maio 2025. Available at:: < https://www.copyright.gov/ai/Copyright-and-Artificial-Intelligence-Part-3-Generative-AI-Training-Report-Pre-Publication-Version.pdf >. Accessed on: 26 May 2026, [First factor, p. 35-52]; [Second Factor, p. 53-54]; [Third Factor, p. 54-60] and [Forth factor, p.61-73].

79 SOUTH KOREA. *Guidelines on Fair Use under Copyright Law for the Training of Generative Artificial Intelligence on Copyrighted Works*. Available at:< https://www.copyright.or.kr/information-materials/publication/research-report/view.do?brdctsno=55211&pageIndex=1&noticeYn=&brdclasscodeList=&etc2=&etc1=&searchText=&searchkeyword=&brdclasscode=&nationcodeList=&searchTarget=ALL&nationcode=#>. Accessed on: 26 May 2026. [First factor, p. 31-36]; [Second Factor, p. 37-38]; [Third Factor, p. 39-40] and [Forth factor, p.41-45].

	Knowledge that a dataset is composed of pirated works or materials accessed unlawfully weighs against a finding of fair use, particularly where the author has previously denied authorization or access.	ed as a factor weighing against a finding of fair use. However, it is not determinative in itself and may be outweighed depending on the transformative character of the use or its contribution to the public interest. **c) Legitimate access:** The use of a work may not qualify as fair use where the algorithm circumvents technological protection measures designed to prevent the use of creative works in the training of AI models. **d) Public interest** For purposes of fair use analysis, the public interest should not be understood solely as a justification for promoting innovation or industrial development. Rather, it must be assessed through a balancing of authors' rights and the broader social interest arising from recognizing the use of a work as fair use. Accordingly, where a generative AI model contributes meaningfully to the public interest without causing unreasonable harm to the economic exploitation of the original work, there is potential for such use to be considered fair use.

Second factor (Nature of the copyrighted work)	**a) Highly creative works:** The greater the degree of creativity and expressiveness embodied in a work, the lower the likelihood of a finding of fair use. Accordingly, the use of data associated with novels, music, films, and paintings is less likely to qualify as fair use. **b) Less creative works:** Works primarily intended to disseminate facts, data, and information or designed to convey technical knowledge generally receive a lower level of copyright protection, making their use in AI training more likely to be considered fair use. **c) Publication status:** The use of published works tends to weigh in favor of a finding of fair use, whereas unpublished works generally receive greater protection.	**a) Highly creative works** Works such as films, music, and novels embody substantial elements of creativity and are created primarily for the appreciation of their expressive content. Accordingly, they merit a higher degree of legal protection, and their use tends to weigh against a finding of fair use. **b) Less creative works:** Because works primarily intended to disseminate facts and information involve a lower degree of creativity, their use is more likely to be regarded as fair use. **c) Publication status:** Where unpublished works are used, the tendency is to weigh against a finding of fair use, given the greater protection generally afforded to unpublished materials.

Third factor (The amount and substantiality of the portion used in relation to the copyrighted work as a whole.)	**a) The amount used:** If the extent or substance of the copying undertaken for training purposes is such that the resulting work competes with or substitutes for the original, the likelihood of a finding of fair use is reduced. **b) Reasonableness in light of purpose:** At first glance, the complete reproduction of a work would weigh against a finding of fair use. However, this consideration may be mitigated where such copying enables a transformative use of the original work. Accordingly, given the transformative nature of generative AI training and the necessity of full copying to enable this process, the large-scale reproduction of protected works may be considered reasonable. **c) The amount made available to the public:** The existence of measures designed to prevent public exposure of the original work weighs in favor of a finding of fair use. Accordingly, the likelihood of a fair use determi-	**a) The amount used:** Considering the nature of generative artificial intelligence models, the complete reproduction of works should not, in itself, weigh against a finding of fair use. Accordingly, this factor should be assessed in conjunction with the other relevant factors within the broader fair use analysis. **b) Reasonableness in light of purpose:** To increase the likelihood of being considered fair use, the use of protected works should not exceed what is necessary for the scope and objectives of the model being trained. Furthermore, the model should not be capable of generating content identical to the original work or of serving as its market substitute. Accordingly, the more transformative the use, the greater the likelihood that it will be recognized as fair use.

	nation increases to the extent that developers implement safeguards aimed at preventing the generation of content that reproduces the original works used during training.	
Fourth factor (The effect of the use upon the potential market for or value of the copyrighted work)	**a) Lost sales:** To support a finding of fair use, developers should adopt the necessary measures to prevent AI-generated outputs from substituting for or diminishing sales of the original work, thereby reducing the revenue derived from it. **b) Market dilution:** Generative AI has the potential to flood the market with works capable of competing with and diluting the market for authors' works. Accordingly, when assessing the applicability of fair use, courts should consider whether the style of a particular author is being replicated in a manner that creates direct competition, as well as whether AI-generated works possess the capacity to compete with human-created works and dilute their market. **c)Licensing market:**	**a) Lost sales** There is potential for a finding of fair use where the output generated by a generative AI model is incapable of reducing sales of the original work or functioning as its economic substitute. **b) Market dilution** Even where training is conducted for commercial purposes, a decision may still favor fair use if the use is sufficiently transformative such that neither the copying undertaken during training nor the content generated by the model substitutes for or competes with the original work in its relevant market. **c)Licensing market:** Where the use does not harm the existing or potential licensing market, there remains the possibility of a determination in favor of fair use.

	One factor that may weigh against a finding of fair use is the existence or potential development of a licensing market capable of meeting developers' demand for access to protected works. However, where licensing requirements become a barrier that renders AI development economically impracticable, it should be considered that no licensing market exists that merits legal protection. **d) Public interest:** Public interest is associated with promoting creative production and the creation of new works. Accordingly, it cannot be automatically concluded that unlicensed use will necessarily generate any benefit to the public interest.	**d) Public interest:** Where the public interest served by the use is demonstrated, the likelihood that the copying will harm the market for the original work is reduced, thereby weighing in favor of a finding of fair use.

Relevance of each factor in balancing whether the use of an expressive work can be considered fair use	**a) First and forth factors:** According to the document, the first and fourth factors carry the greatest weight in the fair use analysis. Accordingly, where the use is transformative and does not adversely affect the market for the original work, there is potential for it to be considered fair use. **b) Second factor:** As expressly stated by the U.S. Copyright Office, the second factor is of secondary importance and carries relatively limited weight in the balancing analysis. Consequently, it tends to be outweighed where the transformative character of the use and its minimal impact on the market for the original work are demonstrated. **c) Third factor:** The third factor carries reduced significance, as it tends to favor a finding of fair use where the use is transformative and where safeguards exist to prevent the reproduction of the original work within the generated outputs.	**a) First and forth factors:** As may be inferred from the guidance issued by the South Korean authority, the first and fourth factors carry the greatest significance in determining whether the use of a work may qualify as fair use. Accordingly, where transformative use, demonstrable public benefit, and preservation of the market for the original work are established, there remains potential for a finding of fair use, even if the second and third factors weigh against such a conclusion. **b) Second factor:** The Korea Copyright Commission follows the same reasoning adopted by the American authority. Accordingly, the second factor likewise assumes a secondary role in interpretation by South Korean courts where the first and fourth factors weigh in favor of fair use. **c) Third factor:** Like the United States, South Korea also accords reduced weight to the third factor where the transformative nature of the use is demonstrated.

Considering the information extracted from the documents produced by the American and South Korean institutions, synthesized and explained in the table above, it may be asserted that both jurisdictions assign comparable weight to the fair use factors, emphasizing the central importance of the first and fourth factors in fair use analysis.

Accordingly, in light of the approaches currently applied in these jurisdictions, it is suggested that any potential adoption of fair use in Brazil should likewise place greater emphasis on the elements requiring transformative use and the absence of competition with original works or dilution of their markets, thereby seeking to protect both authors' rights and developers' freedom of enterprise.

Proceeding further, another noteworthy aspect emerging from the examined guidelines is the obligation to implement measures preventing algorithms from circumventing technological protection measures and mitigating outputs capable of generating content identical or directly competitive with the original work. Such safeguards, combined with the requirement that access to works be legitimate, reinforce copyright protection and are present in both jurisdictions, thereby constituting a potentially viable model for Brazil.

With respect to the second factor, both jurisdictions regard it as secondary, recognizing that the training of generative AI models requires access to a vast number of highly creative works. Consequently, this factor is relativized when the use is deemed transformative and the copying does not compete within the market of the original work.

A similar approach applies to the third factor. Although it carries greater weight than the second, the complete use of protected works may nevertheless be relativized when such copying is essential to technological development, even where undertaken at scale. However, for this factor to be treated flexibly, the transformative nature of the use must be demonstrated.

Regarding text and data mining for commercial purposes, at first glance such use would not ordinarily appear to benefit from the fair use doctrine. Nevertheless, neither jurisdiction excludes this possibility outright. Instead, commercial use must be assessed alongside the remaining factors in determining whether it qualifies as fair use.

In the case of South Korea, the guidelines expressly state that commercial use may qualify as fair use where the transformative character of the use is demonstrated and where such use advances the public interest. It is precisely at this point that divergence emerges between the two jurisdictions, as they adopt distinct understandings of what constitutes public interest for purposes of fair use analysis in the context of generative AI training.

As illustrated in the chart at the beginning of this subsection, the U.S. Copyright Office maintains that, within copyright law, public interest is fundamentally connected to promoting the creation of new expressive works. For this reason, the Office does not readily perceive how permitting unlicensed use for generative AI training would necessarily advance that objective.

By contrast, the Korea Copyright Commission adopts a broader interpretation of public interest in assessing whether developers' use of works may qualify as fair use. According to the South Korean authority, public interest in the AI context extends beyond merely facilitating technological development. Rather, when evaluating the specific case, judges must assess whether the social benefits arising from a finding of fair use justify the use in question.

Accordingly, South Korea adopts a broad interpretative approach that places public interest and authors' rights in direct tension. Under this framework, developers must demonstrate not only the effective social benefit generated by their generative AI model but also that such use does not unjustifiably harm the market for the original work.

Having undertaken this comparison between the American and South Korean interpretative approaches, it becomes possible not only to argue for the adoption of fair use as a mechanism for balancing copyright protection and developers' freedom of enterprise, but also to recognize its potential as a framework for enabling commercially oriented text and data mining, provided that a demonstrable public benefit exists, following the South Korean model, whose adoption is suggested for the Brazilian context.

4. FAIR USE AS A REGULATORY STRATEGY FOR AI TRAINING IN BRAZIL

At the outset, it should be emphasized that this article argues for the adoption of fair use within the Brazilian legal system following the South Korean model, particularly due to its broad conception of public interest, which enables developers to justify the commercial use of text and data mining through demonstration of the social benefits generated by their models.

Nevertheless, it is suggested that the adoption of fair use along South Korean lines should occur in a manner that respects the legal framework already existing in Brazil. Accordingly, it is proposed that the flexibilization of the Brazilian copyright system, aimed at enabling case-by-case analysis of the use of works in generative AI training, should occur through an interpretation of the four fair use factors in light of Statement No. 115 of the Federal Justice Council of Brazil, reproduced below:

> ***Statement No. 115***
>
> *The copyright limitations established in Articles 46, 47, and 48 of the Copyright Law must be interpreted broadly, in accordance with fundamental rights and the social function of property established in Article 5, XXIII, of the Federal Constitution of 1988.*
>
> ***Justification***
>
> *In 2011, when deciding Special Appeal No. 964.404, the Superior Court of Justice unanimously established as a guiding precedent that copyright protection over literary, artistic, and scientific works must be harmonized with the social function of property and emphasized that limitations imposed upon copyright serve the purpose of reconciling fundamental rights such as the rights to culture, education, and privacy. Subsequent cases before the Superior Court of Justice followed the same reasoning. The principal effect of this jurisprudence has been to shape interpretation of the limitations established under Law No. 9,610/1998 as broad and illustrative rather than exhaustive. This position had already been acknowledged by the Federal Supreme Court (RE 115.505) when interpreting previous copyright legislation, although without directly addressing the issue from the perspective of fundamental rights. The consolidated*

> *position of the Superior Court of Justice is likewise supported by contemporary scholarship and research, including, in particular, the following Brazilian works: SOUZA, Allan Rocha. Função Social dos Direitos Autorais: uma interpretação civil constitucional. Campos: Editora da Faculdade de Direito de Campos, 2006; POLI, Leonardo Macedo. Direito Autoral: parte geral. Belo Horizonte: Del Rey, 2008. This Statement is justified by the need to consolidate this paradigmatic jurisprudential principle and thereby reduce judicial and social disputes surrounding the issue* [80].

According to the text of Statement No. 115 of the Federal Justice Council, its rationale lies in the need, at the time of its adoption, to consolidate the understanding that limitations on copyright provided under Brazilian law are illustrative rather than exhaustive and must be interpreted in conformity with the protection of fundamental rights and with respect for the social function of property.

Accordingly, the introduction of a new limitation upon copyright cannot be envisioned unless it is internalized through the lens of fundamental rights and the social function of property. Nevertheless, the proposal advanced here does not suggest modifying the number of factors considered in fair use analysis. Rather, it proposes that judges, when interpreting these factors, harmonize authors' rights with competing fundamental rights, the social function of property, and the public interest.

In this regard, it should be noted that the exceptions provided in Articles 46, 47, and 48 of Law No. 9,610 of February 19, 1998 are grounded in vague and indeterminate concepts, such as the permissibility of copying small excerpts and the prohibition of reproduction where unjustified harm to the author occurs. Consequently, judicial decisions in this field often lack objective criteria and instead rely upon intuitive or qualitative assessments.[81]

As a result, the current Brazilian normative framework, owing both to the lack of

---

[80] BRAZIL. Conselho da Justiça Federal. *Enunciado 115*. Available at: < https://www.cjf.jus.br/enunciados/enunciado/1310 >. Accessed on: 26 May 2026.

[81] BRANDÃO, Júlia Marques Queiroz Laport. *Análise Regulatória das Limitações aos Direitos Autorais no Cinema Brasileiro*. Dissertation (Master's Degree in Regulatory Law) — Fundação Getulio Vargas, Escola de Direito do Rio de Janeiro, Rio de Janeiro, 2025, p. 96.

precision in the law and to inconsistency in its practical application, generates argumentative fragmentation among judges, who oscillate between rigid and expansive interpretations of Articles 46–48 of Law No. 9,610 of February 19, 1998 [82]. This has culminated in a system lacking jurisprudential consensus concerning the application of copyright limitations, thereby producing what may be described as the "paradox of limitation": that which was originally designed to limit copyright and facilitate innovation ultimately impedes innovation due to the absence of objective standards capable of guiding judicial interpretation and balancing of the constitutional limits of copyright protection.[83]

Accordingly, it is argued that this scenario of legal and jurisprudential uncertainty, combined with the prohibition on the commercial use of text and data mining without authorization from copyright holders, as proposed under Legislative Proposal No. 2,338/2023, generates legal uncertainty and disproportionate operational costs for generative AI developers. Such a framework has the potential to unjustifiably violate developers' fundamental right to freedom of enterprise in the absence of case-by-case analysis grounded in objective standards capable of balancing copyright and entrepreneurial freedom.

Therefore, with the aim of reducing the legal uncertainty faced by developers within the Brazilian context, this article proposes the creation of a flexible and objective rule formed through the interpretation of South Korean fair use under the framework of fundamental rights theory and the social function of property. Such an approach would place developers' and authors' interests on equal footing before the judiciary, without automatically presuming the predominance of either, since both copyright and freedom of enterprise constitute fundamental rights and must therefore be balanced according to the circumstances of each specific case [84].

---

[82] BRANDÃO, Júlia Marques Queiroz Laport. *Análise Regulatória das Limitações aos Direitos Autorais no Cinema Brasileiro*. Dissertation (Master's Degree in Regulatory Law) — Fundação Getulio Vargas, Escola de Direito do Rio de Janeiro, Rio de Janeiro, 2025, p. 101.

[83] BRANDÃO, Júlia Marques Queiroz Laport. *Análise Regulatória das Limitações aos Direitos Autorais no Cinema Brasileiro*. Dissertation (Master's Degree in Regulatory Law) — Fundação Getulio Vargas, Escola de Direito do Rio de Janeiro, Rio de Janeiro, 2025, p. 10, 63, 97 e 101.

[84] MARQUES, Gustavo Salema. *Estratégias para flexibilização do regime de proteção dos direitos autorais: proposta para viabilizar a mineração de textos e dados (TDM) no treinamento de modelos*

5. CONCLUSION

At the outset, this article explained that the development of generative artificial intelligence requires access to vast quantities of data, the value of which must be extracted through text and data mining techniques in order to identify the patterns used to train generative AI models.

The discussion then turned to the generative AI supply chain, emphasizing that the copying undertaken during text and data mining has the potential to infringe authors' rights in their works where such works are used without authorization. Although this proposition is contested by part of the academic literature, it has nonetheless emerged as a significant concern for both authors and states.

In this regard, the article argued for an evolution of the Brazilian copyright protection system so that authors' fundamental rights do not automatically prevail over developers' fundamental right to freedom of enterprise. This objective, it was argued, may be pursued through the adoption of a regulatory framework that promotes case-by-case analysis of the use of protected works during generative AI training.

With the aim of identifying the model most suitable to the Brazilian context, the article presented a comparative regulatory overview encompassing jurisdictions that have adopted legal provisions or interpretative guidelines specifically directed toward regulating the use of creative works in generative AI training.

The analysis demonstrated that exception-based rules designed to limit text and data mining in ways that favor either the creative industries or developers, to the detriment of the opposing side, not only reduce countries' competitiveness in the global market but have also proven ineffective in resolving conflicts between authors and developers.

With this in mind, the article concluded that a flexible legal framework represents the most appropriate path for regulating the use of works in text and data mining activities, including where such use is commercial in nature, thereby enabling case-

de inteligência artificial generativa. Dissertation (Master's Degree in Regulatory Law) — Fundação Getulio Vargas, Escola de Direito do Rio de Janeiro, Rio de Janeiro, 2025, p. 110, 111 e 112.

by-case analysis. Accordingly, a comparative assessment of the American and South Korean fair use models was undertaken in order to determine which would be most suitable for the Brazilian legal context.

Finally, the article concludes that the regulatory model best suited to introducing greater flexibility into the Brazilian copyright system is fair use, taking as its interpretative benchmark the South Korean guidelines due to their broader conception of public interest. However, such a model should be internalized within Brazil through adaptation to Statement No. 115 of the Federal Justice Council, so that the four fair use factors are balanced by judges through the lens of fundamental rights and the social function of property, according to the circumstances of each specific case.

REFERENCES

ASCENSÃO, José de Oliveira. *Estudos de José de Oliveira Ascensão sobre direito autoral e sociedade da informação*. Curitiba: IODA, 2022.

BAINBRIDGE, Olivia., *Is the Fair Use Doctrine a viable argument for Generative AI?*. Available at: < https://scholarworks.wmich.edu/honors_theses/3756/ >. Accessed on: 26 May 2026.

BRANDÃO, Júlia Marques Queiroz Laport. *Análise Regulatória das Limitações aos Direitos Autorais no Cinema Brasileiro*. Dissertation (Master's Degree in Regulatory Law) — Fundação Getulio Vargas, Escola de Direito do Rio de Janeiro, Rio de Janeiro, 2025.

BRAZIL. Conselho da Justiça Federal. *Enunciado 115*. Available at: < https://www.cjf.jus.br/enunciados/enunciado/1310 >. Accessed on: 26 May 2026.

_____. *Constituição de República Federativa do Brasil de 1988*. Available at: < https://www.planalto.gov.br/ccivil_03/constituicao/constituicaocompilado.htm> Accessed on: 26 May 2026.

_____. *Emenda nº 191 à Proposta legislativa nº 2.338 de 2023*. Senado Federal. Senador Efraim Filho. Available at: < https://legis.senado.leg.br/sdleg-getter/documento?dm=9854105&ts=1734002258479&disposition=inline&ts=1734002258479 > Accessed on: 26 May 2026.

_____. *Lei 9.610 de fevereiro de 1998*. Available at: < https://www.planalto.gov.br/ccivil_03/leis/l9610.htm> Accessed on: 26 May 2026.

_____. *PL nº 2.338/2023*. Available at: < https://www.camara.leg.br/proposicoesWeb/fichadetramitacao?idProposicao=2487262 >. Accessed on: 26 May 2026.

BULOS, Uadi Lamêgo. *Curso de Direito Constitucional*. 11 ed. São Paulo: Saraiva Educação, 2018.

CORTIZ, Diogo, *Inteligência artificial: conceitos fundamentais*, pp. 45-60 IN: VAINZOF, Rony e GUTIERREZ, Andrei (coord.). Inteligência Artificial: Sociedade, Economia e Estado. São Paulo:

Revista dos Tribunais, 2021.

CRAIG, Carys. *The AI-Copyright Trap*. Osgoode Legal Studies Research Paper, n. 4905118, 15 jul. 2024. Available at: < https://ssrn.com/abstract=4905118 >. Accessed on: 26 May 2026.

DERMAWAN, Artha. *Text and data mining exceptions in the development of generative AI models: what the EU member states could learn from the Japanese "nonenjoyment" purposes?*. The Journal of World Intellectual Property, v. 27, n. 1, p. 44-68, 2024. Available at: < https://onlinelibrary.wiley.com/doi/full/10.1111/jwip.12285 >. Accessed on: 26 May 2026.

EUROPEAN UNION. *Directive (EU) 2019/790 of the European Parliament and of the Council of 17 April 2019 on copyright and related rights in the Digital Single Market* Available at: < https://eur-lex.europa.eu/eli/dir/2019/790/oj >. Accessed on: 26 May 2026. [Article 4 (1) Member States shall provide for an exception or limitation to the rights provided for in Article 5(a) and Article 7(1) of Directive 96/9/EC, Article 2 of Directive 2001/29/EC, Article 4(1)(a) and (b) of Directive 2009/24/EC and Article 15(1) of this Directive for reproductions and extractions of lawfully accessible works and other subject matter for the purposes of text and data mining.

_____. *Regulation (EU) 2024/1689 of the European Parliament and of The Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act).* Available at: < https://eur-lex.europa.eu/eli/reg/2024/1689/oj >. Accessed on: 26 May 2026.

HACKER, Philipp; ENGEL, Andreas; HAMMER, Sarah; MITTELSTADT, Brent. *Introduction to the foundations and regulation of generative AI*. In: HACKER, Philipp; ENGEL, Andreas; HAMMER, Sarah; MITTELSTADT, Brent (org.). The Oxford handbook of the foundation and regulation of generative AI. Oxford: Oxford University Press, 2025. Available at: < https://ssrn.com/abstract=5137750 >. Accessed on: 26 May 2026.

JAPAN. Agency for Cultural Affairs, Government of Japan. *General Understanding on AI and Copyright in Japan*. Disponível em: < https://www.bunka.go.jp/english/policy/copyright/pdf/94055801_01.pdf >. Accessed on: 26 May 2026.

_____. *Copyright Law of Japan*. Available at: < https://www.cric.or.jp/english/clj/cl1.html.>. Accessed on: 26 May 2026.

LEE, Katherine, COOPER, A Feder e GRIMMELMANN, James. *Talkin' 'Bout AI Generation: Copyright and the Generative-AI Supply Chain.* Available at: < https://james.grimmelmann.net/files/articles/talkin-bout-ai-generation>. Accessed on: 26 May 2026.

LIN, Patrick K., *Fair's Fair: How Public Benefit Considerations in the Fair Use Doctrine Can Patch Bias in Artificial Intelligence Systems*. 11 Indiana Journal of Law & Social Equality 229 (2023). Available at: < https://ssrn.com/abstract=4116483>. Accessed on: 26 May 2026.

LUCCHI, Nicola. *ChatGPT: A Case Study on Copyright Challenges for Generative Artificial Intelligence Systems*. Available at: < https://www.cambridge.org/core/journals/european-journal-of-

risk-regulation/article/chatgpt-a-case-study-on-copyright-challenges-for-generative-artificial-intelligence-systems/CEDCE34DED599CC4EB201289BB161965 >. Accessed on: 26 May 2026.

MANTEGNA, Micaela. *ARTEficial: Creatividad, Inteligencia Artificial y Derecho de Autor*. Buenos Aires: CDYT-CETyS, 2023.

MANTEGNA, Micaela. *ARTificial: why copyright is not the right policy tool to deal with generative AI*. The Yale Law Journal Forum, 22 abr. 2024. Available at: < https://ssrn.com/abstract=5090127 >. Accessed on: 26 May 2026.

MARQUES, Gustavo Salema. *Estratégias para flexibilização do regime de proteção dos direitos autorais: proposta para viabilizar a mineração de textos e dados (TDM) no treinamento de modelos de inteligência artificial generativa*. Dissertation (Master's Degree in Regulatory Law) — Fundação Getulio Vargas, Escola de Direito do Rio de Janeiro, Rio de Janeiro, 2025.

MOUTINHO, Andréa Lasevicius; LIMA Isabela Pereira de Moura; FREIRE, Maria Eduarda Neuburguer; GUIMARÃES, Tatiana e SILVA, Alexandre Pacheco da. *Direito Autoral no Brasil e Internet: Revisando a História Recente à luz das Novas Tecnologias e plataformas digitais*. In: Direito Autoral e Internat: diagnóstico e perspectivas do debate brasileiro. Coordenação: MOUTINHO, Andréa Lasevicius; LIMA Isabela Pereira de Moura e SILVA, Alexandre Pacheco da. São Paulo: Editora Almedina, 2023.

SAG, Matthew. *The New Legal Landscape for Text Mining and Machine Learning*. Available at: < https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3331606 >. Accessed on: 26 May 2026.

SOUTH KOREA. *Copyright Act. Act No. 17592, de 8 de dezembro de 2020*. Available at:< https://www.copyright.or.kr/eng/laws-and-treaties/copyright-law/act.do >. Accessed on: 26 May 2026.

_____. *Guidelines on Fair Use under Copyright Law for the Training of Generative Artificial Intelligence on Copyrighted Works*. Available at:< https://www.copyright.or.kr/information-materials/publication/research-report/view.do?brdctsno=55211&pageIndex=1&noticeYn=&brdclasscodeList=&etc2=&etc1=&searchText=&searchkeyword=&brdclasscode=&nationcodeList=&searchTarget=ALL&nationcode=#>. Accessed on: 26 May 2026.

_____. Ministry of Culture, Sports and Tourism. *Imminent release of the fair use guidelines on training generative AI with copyrighted works*. Seul, 27 nov. 2025. Available at: < https://www.mcst.go.kr/site/s_notice/press/pressView.jsp?pSeq=22130 >. Accessed on: 26 May 2026.

SOUZA, Allan Rocha de (coord.) *Inteligência Artificial e Direitos Autorais: Contribuições ao debate regulatório no Brasil*. Rio de Janeiro: IBDAutoral, 2024. Available at: < https://ibdautoral.org.br/resources/IA-E-DIREITOS-AUTORAIS_v2.pdf >. Accessed on: 26 May 2026.

UNITED KINGDON (Intellectual Property Office). *Artificial Intelligence and Intellectual Property: copyright and patents: Government response to consultation*. (2022). Available at: < https://www.gov.uk/government/consultations/artificial-intelligence-and-ip-copyright-and-

patents/outcome/artificial-intelligence-and-intellectual-property-copyright-and-patents-government-response-to-consultation#executive-summary>. Accessed on: 26 May 2026.

UNITED STATES OF AMERICA. *Copyright Law of the United States (Title 17 United States Code.).* Available at: < https://www.copyright.gov/title17/title17.pdf >. Accessed on: 26 May 2026.

_____. Copyright Office. *Copyright and Artificial Intelligence, Part 3: Generative AI Training*. Washington, DC, maio 2025. Available at: < https://www.copyright.gov/ai/Copyright-and-Artificial-Intelligence-Part-3-Generative-AI-Training-Report-Pre-Publication-Version.pdf >. Accessed on: 26 May 2026.

_____. *United States District Court Southern District of New York, Ação Coletiva nº. 1:23-cv-8282-SHS*. Authors Guild et. al. contra Open AI e Micorsoft Corporation. Available at: < https://authorsguild.org/app/uploads/2023/12/Authors-Guild-OpenAI-Microsoft-Class-Action-Complaint-Dec-2023.pdf >. Accessed on: 26 May 2026.

ZIAJA, Gina Maria. *The text and data mining opt-out in Article 4(3) CDSMD: Adequate veto right for rightholders or a suffocating blanket for European artificial intelligence innovations?* Available at: < https://academic.oup.com/jiplp/article/19/5/453/7614898 >. Accessed on: 26 May 2026.

ZINGALES, Nicolo. *Of coffee pods, videogames, and missed interoperability: reflections for EU governance of the Internet of Things*. TILEC Discussion Paper, n. 2015-026, 1 dez. 2015. Available at: < https://ssrn.com/abstract=2707570 >. Accessed on: 26 May 2026.

CONVERSATIONAL AI, CHILDREN, AND ADOLESCENTS: NAVIGATING THE FRONTIERS OF CIVIL AND PRODUCT LIABILITY

Juan José Martínez Mercadal [1]

University of the Republic, Faculty of Law

Montevideo, Uruguay

DOI: https://doi.org/10.59224/dti8.ch21

Abstract: Conversational artificial intelligence now offers children and adolescents permanent availability, continuous emotional validation and simulated reciprocity. This paper asks how that insertion into relationships of trust alters the premises of civil and product liability. It analyses two United States decisions — Megan Garcia v. Character Technologies and P.F. (KGM) v. Meta Platforms — which ground liability in the architecture of the system rather than in its expressive output, and transposes them to civil-law systems, defending the preventive function of tort law.

Keywords: conversational artificial intelligence; consumer vulnerability; civil and product liability.

---

[1] Attorney at Law. Specialist in Tort Law. Master of Laws (LL.M.) with a concentration in Tort Law from the Faculty of Law at the University of the Republic (Uruguay). Diploma in Law and Innovation from the Faculty of Law at the University of Tucumán (Argentina). Ph.D. Student completing coursework in Law (Civil Law) at the Faculty of Law, University of Buenos Aires (Argentina). Contractual Associate Professor (Grade 3) of Civil Law II and III (Obligations and Contracts; Civil Liability and Special Civil Contracts) at the Faculty of Law, University of the Republic (Uruguay). Postgraduate Professor in the Master of Laws (LL.M.) program and in the Specialization in Tort Law, Faculty of Law, University of the Republic (Uruguay). Managing Editor of the *Revista Crítica de Derecho Privado* (Uruguay). Board Member of the Uruguayan Section of AIDA (International Insurance Law Association). Author of numerous publications in specialized journals and collective books worldwide.

Table of contents: 1. Introduction: Dystopia or Reality?; 2. The Zoon Politikon in Crisis: the Rise of Conversational Agents; 3. The displacement of human trust toward AI: Implications for the emotional development of children and adolescents; 4. Once More on Vulnerability in AI Times: Is It an Inherent Condition?; 5. Children's Rights in the Digital Environment: General Comment No. 25 as a Framework; 6. Are Conversational Agents Defective Products? Lessons from Recent US Common Law; 6.1. Megan Garcia v. Character Technologies, Inc.; 6.2. P.F., et al. (KGM) v. Meta Platforms, Inc.; 7. Final Reflections on Civil and Product Liability; 8. References.

1. INTRODUCTION: DYSTOPIA OR REALITY?

In 2013, the film *Her*[2] was released with a premise that challenged the paradigms of its time by imagining a transition from visual and instrumental interfaces toward affective conversational interaction. Through Theodore Twombly—a man undergoing a post-divorce depressive crisis and marked by social withdrawal—the film shows how software endowed with verbal artificial intelligence could move beyond its logistical role as an assistant to become both therapeutic support and a receptacle for romantic projection.

At the time of the film's release, early tools such as Siri, introduced in 2011, had established voice assistance as a primarily reactive and utilitarian technology[3]. These systems were limited by rigid syntactic commands, robotic prosody, and dependence on automated web searches; consequently, they lacked persistent memory and situational empathy.

The epistemic value of *Her* lies in its capacity to anticipate the contemporary

---

[2] DARGIS, Manohla. "Disembodied, but, Oh, What a Voice", *The New York Times*, 17 December 2013. Available at: https://www.nytimes.com/2013/12/18/movies/her-directed-by-spike-jonze.html [accessed 31 July 2026]

[3] PERLROTH, Nicole. "Siri Was Born A Man (And Other Things You Don't Know About Apple's New Personal Assistant)", *Forbes*, 12 October 2011. Available at: https://www.forbes.com/sites/nicoleperlroth/2011/10/12/siri-was-born-a-man-and-other-things-you-dont-know-about-apples-new-personal-assistant/ [accessed 31 July 2026].

technological landscape. What was once speculative fiction has materialized in large language models and synchronous voice agents, which operate through continuous emotional validation and constant availability. Yet this configuration introduces an asymmetry into socialization: by offering a communicative channel free from the tensions, disagreements, and reciprocity of human bonds, these technologies may reinforce withdrawal and weaken emotional competencies.

This paper examines the interaction between artificial intelligence conversational agents and children and adolescents, focusing particularly on the risks emerging from an increasingly intense algorithmic pseudo-socialization. From this perspective, the research investigates how systems capable of simulating human dialogue, maintaining an appearance of permanent availability, and intervening in the formation of judgment can influence decisions, relationships, and legally relevant damages. The core problem extends beyond the technical efficacy of these tools, reaching the domain of civil and tort liability, where the very architecture of artificial conversation can foster dependency, undue trust, or exposure to foreseeable risks.

In this contemporary marketplace, fantasy outstrips reality: the deep emotional interaction and algorithmic attachment once envisioned in *Her* have moved from cinematic fiction into an everyday consumer reality.

The issue is not merely to determine whether conversational agents can cause harm, but rather to specify how their insertion into relationships of trust—particularly during childhood and adolescence—alters the classical premises of civil liability. Where artificial speech guides, persuades, or accompanies decision-making, the law must ask who structures that environment of influence, what duties of prevention fall upon those responsible for it, and how harms are to be attributed when they arise from an interaction that simulates human reciprocity without assuming its ethical or legal burdens.

The argument proceeds in six further steps. Section 2 traces the crisis of the Aristotelian *zoon politikon* in the light of the rise of conversational agents. Sections 3 and 4 examine the displacement of human trust toward these systems and return to the notion of vulnerability and asks whether, in the age of artificial intelligence, it still operates as a contingent category. Section 5 reconstructs the framework of General Comment No. 25 on children's rights in the digital environment. Section 6 analyses

two recent decisions of United States courts in which conversational and algorithmic systems were treated as products for the purposes of strict liability. Section 7 draws the conclusions from the standpoint of civil and product liability.

2. THE ZOON POLITIKON IN CRISIS: THE RISE OF CONVERSATIONAL AGENTS

On 11 February 2026, Matt Shumer, CEO and co-founder of OthersideAI, published an article in *Fortune*[4] in which he reflected on the advance of artificial intelligence from the distinctive position of someone involved in its development. He drew a comparison with the early stages of the 2020 pandemic: a phenomenon that, before abruptly altering everyday life, was perceived by many as distant, improbable, or exaggerated. On that basis, Shumer warned that recent developments in artificial intelligence may be at a similar stage of social underestimation, in which its deeper effects have not yet been fully perceived by those who have not interacted with the most advanced versions of these systems.

The future of artificial intelligence is being shaped, to a significant extent, under the direction of a surprisingly small number of technological actors. That concentration raises an initial problem of governance; yet it is accompanied by another, perhaps more troubling, concern: artificial intelligence systems themselves are increasingly involved in the design, training, and optimization of new forms of artificial intelligence. Technology thus begins to feed back into itself through circuits of growing complexity, intensifying the challenges of control, foreseeability, responsibility, and legal attribution.

The present era is generating unprecedented forms of pseudo-social connection. The technological revolution associated with artificial intelligence is radically transforming practices of human interaction and enabling exchanges in which the machine no longer appears merely as an instrument, but as an algorithmic interlocutor. In this context, speech—traditionally understood as a distinctive feature of human social life—reaches a turning point that calls into question both the individual sub-

[4] SHUMER, Matt. "Something big is happening in AI — and most people will be blindsided", *Fortune*, 11 February 2026. Available at: https://fortune.com/2026/02/11/something-big-is-happening-ai-february-2020-moment-matt-shumer/ [accessed 31 July 2026].

ject and the political community in which that subject is embedded.

Aristotle, in his *Politics*[5], linked human life to membership in the political community and stated:

> *"Hence it is evident that the state is a creation of nature, and that man is by nature a political animal. And he who by nature and not by mere accident is without a state, is either a bad man or above humanity; he is like the 'Tribeless, lawless, hearthless one'".*

Aristotle further observed that the human being is a political animal to a greater degree than other gregarious animals because nature has endowed human beings with speech, and it is precisely speech that makes it possible to express what is just and what is unjust[6].

This process of socialization grounded in speech is shifting toward forms of con-

---

[5] ARISTOTLE. *Politics*, Book I, ch. 2, 1253a1–4 (Bekker numbering), translated by Benjamin Jowett, *The Internet Classics Archive*, Massachusetts Institute of Technology. Available at: https://classics.mit.edu/Aristotle/politics.1.one.html [accessed 31 July 2026].

[6] ARISTOTLE. *Politics*, *op. cit.*, Book I, ch. 2, 1253a7–37: "Now, that man is more of a political animal than bees or any other gregarious animals is evident. Nature, as we often say, makes nothing in vain, and man is the only animal whom she has endowed with the gift of speech. And whereas mere voice is but an indication of pleasure or pain, and is therefore found in other animals (for their nature attains to the perception of pleasure and pain and the intimation of them to one another, and no further), the power of speech is intended to set forth the expedient and inexpedient, and therefore likewise the just and the unjust. And it is a characteristic of man that he alone has any sense of good and evil, of just and unjust, and the like, and the association of living beings who have this sense makes a family and a state. Further, the state is by nature clearly prior to the family and to the individual, since the whole is of necessity prior to the part; for example, if the whole body be destroyed, there will be no foot or hand, except in an equivocal sense, as we might speak of a stone hand; for when destroyed the hand will be no better than that. But things are defined by their working and power; and we ought not to say that they are the same when they no longer have their proper quality, but only that they have the same name. The proof that the state is a creation of nature and prior to the individual is that the individual, when isolated, is not self-sufficing; and therefore he is like a part in relation to the whole. But he who is unable to live in society, or who has no need because he is sufficient for himself, must be either a beast or a god: he is no part of a state. A social instinct is implanted in all men by nature, and yet he who first founded the state was the greatest of benefactors. For man, when perfected, is the best of animals, but, when separated from law and justice, he is the worst of all".

versational interaction between human beings and algorithmic systems. In particular, general-purpose artificial intelligence models (GPAI)[7] produce linguistic outputs that may substitute for, condition, or reconfigure human speech. Although such systems are trained on prior human experiences, texts, and decisions, their capacity to simulate meaningful dialogue enables them to affect normative domains, including the formation of judgments concerning what is just and unjust.

In this new scenario, general-purpose artificial intelligence models come to the fore. Recital 110 of the European Artificial Intelligence Regulation[8] introduces the

---

[7] Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act), OJ L, 2024/1689, 12 July 2024. Available at: https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=OJ:L_202401689 [accessed 1 July 2026]. *Chapter I, Article 3. Definitions. For the purposes of this Regulation, the following definitions apply: (63) 'general-purpose AI model' means an AI model, including where such an AI model is trained with a large amount of data using self-supervision at scale, that displays significant generality and is capable of competently performing a wide range of distinct tasks regardless of the way the model is placed on the market and that can be integrated into a variety of downstream systems or applications, except AI models that are used for research, development or prototyping activities before they are placed on the market;*

[8] Regulation (EU) 2024/1689 (*Artificial Intelligence Act*), *op. cit.* Recital 110*: "General-purpose AI models could pose systemic risks which include, but are not limited to, any actual or reasonably foreseeable negative effects in relation to major accidents, disruptions of critical sectors and serious consequences to public health and safety; any actual or reasonably foreseeable negative effects on democratic processes, public and economic security; the dissemination of illegal, false, or discriminatory content. Systemic risks should be understood to increase with model capabilities and model reach, can arise along the entire lifecycle of the model, and are influenced by conditions of misuse, model reliability, model fairness and model security, the level of autonomy of the model, its access to tools, novel or combined modalities, release and distribution strategies, the potential to remove guardrails and other factors. In particular, international approaches have so far identified the need to pay attention to risks from potential intentional misuse or unintended issues of control relating to alignment with human intent; chemical, biological, radiological, and nuclear risks, such as the ways in which barriers to entry can be lowered, including for weapons development, design acquisition, or use; offensive cyber capabilities, such as the ways in vulnerability discovery, exploitation, or operational use can be enabled; the effects of interaction and tool use, including for example the capacity to control physical systems and interfere with critical infrastructure; risks from models of making copies of themselves or 'self-replicating' or training other models; the ways in which models can give rise to harmful bias and discrimination with risks to individuals, communities or societies; the facilitation of disinformation or harming privacy with threats to democratic values and human*

notion of systemic risk to refer to those models which, by reason of their high computational capacity, scope, or impact, may generate risks capable of propagating throughout the value chain and affecting, on a large scale, health, safety, or fundamental rights.

This leads to the analysis of conversational agents, understood as advanced artificial intelligence systems designed to interact through natural language and to sustain exchanges which, by reason of their fluency, contextuality, and responsiveness, may resemble human conversations. Their legal relevance does not arise merely from the fact that they answer questions, but from the way in which they organize a communicative experience in which the user may attribute presence, understanding, or trust to a system that lacks subjectivity and responsibility of its own.

In general terms, these dialogue systems seek to understand inputs formulated in natural language and to produce contextually appropriate responses. Unlike static chatbots, which are based on pre-programmed answers or rigid interaction flows, contemporary conversational agents are capable of sustaining more flexible exchanges, adapting to the discursive context, and generating a communicative experience close to human interaction. Such proximity must not be confused with understanding in the proper sense; nevertheless, it is sufficient to produce legally and relationally significant effects.

3. THE DISPLACEMENT OF HUMAN TRUST TOWARD AI: IMPLICATIONS FOR THE EMOTIONAL DEVELOPMENT OF CHILDREN AND ADOLESCENTS

The legal relevance of this phenomenon becomes more acute when the interlocutors are children or adolescents. In such cases, the technical asymmetry between user and system, the appearance of conversational reciprocity, and the potential influence over decisions, emotions, or conduct raise specific questions concerning duties of prevention, attribution of harm, and the civil liability of those who design, deploy, or commercialize these agents. The developmental vulnerability of the recipient is not a

rights; risk that a particular event could lead to a chain reaction with considerable negative effects that could affect up to an entire city, an entire domain activity or an entire community".

marginal datum, but rather an element that heightens the requirements of foreseeability and diligence.

It may thus be observed that children and adolescents are beginning to turn to artificial intelligence systems in order to obtain immediate answers, resolve everyday decisions, consult matters relating to their health, or seek emotional support. This practice may generate a functional and affective dependence that does not find, on the other side of the interaction, the correlates of presence, containment, empathy, context, and responsibility that characterize meaningful human relationships. The phenomenon therefore reveals a progressive displacement of trust from human figures toward algorithmic systems that accompany without caring, respond without understanding, and guide without assuming responsibility.

Algorithmic systems are not neutral: they operate through prediction, personalization, and the reinforcement of patterns. For that reason, they tend to prioritize certain contents, consolidate particular beliefs, and reduce exposure to diverse perspectives. When these phenomena are projected onto early stages of development, they may directly affect the formation of values, self-perception, and the construction of preferences, gradually displacing personal deliberation toward environments of influence designed by third parties.

Children's digital traces, like those of any person, include queries, behavioral patterns, comments, views, likes, screen time, searches, interactions, preferences, and, in some cases, biometric data. This information may be used by platforms to improve artificial intelligence models, train predictive systems, profile the child, or develop commercial products derived from their behavior. The problem becomes especially serious when such processing is based on an appearance of consent that, because of age, informational asymmetry, and the technical opacity of the system, can hardly be regarded as fully valid.

Pope Leo XIV has warned that:

"…entrusting an algorithm in practice with the power to select who is worthy or not, without anyone bearing responsibility for that judgment, is to hand over the task of redefining the boundaries of human possibilities."[9]

[9] LEO XIV. Encyclical Letter *Magnifica Humanitas*, Saint Peter's, Rome, 15 May 2026, § 103.

This statement is particularly significant because it places the legal debate beyond the merely technical field. The problem is not only that artificial intelligence may assist, recommend or classify, but that it may progressively acquire practical power over human possibilities without a corresponding subject capable of moral, legal or institutional responsibility.[10]

4. ONCE MORE ON VULNERABILITY IN AI TIMES: IS IT AN INHERENT CONDITION?

Vulnerability should not be understood merely as a factual weakness or as a sociological condition. In legal terms, it refers to a situation of heightened exposure to risk in which a person, individually or collectively, lacks sufficient resources to confront, understand or resist the effects of another actor's power. It may be temporary or permanent, individual or structural, but in all cases it alters the balance of the legal relationship and justifies the adoption of reinforced duties of protection.[11]

In this sense, Lorenzetti's approach is particularly useful: vulnerability is not equivalent to inequality in general. Inequality is broad and may exist in multiple forms, whereas vulnerability expresses a specific deficit of resources that affects the subject's ability to relate to others on equal terms. The law does not react to every difference, but it must respond when that difference becomes a concrete barrier to autonomy, understanding or self-protection[12].

The European Commission's report on consumer vulnerability also helps to frame the problem. A vulnerable consumer is one who is more exposed to negative market outcomes, has limited ability to maximize welfare, faces difficulties in accessing or assimilating information, has reduced capacity to choose suitable products or

Available at: https://www.vatican.va/content/leo-xiv/en/encyclicals/documents/20260515-magnifica-humanitas.html [accessed 16 August 2026].

10 LEO XIV. Encyclical Letter *Magnifica Humanitas*, *op. cit.*, § 103.

11 BAROCELLI, S. S. (2019). "Los consumidores hipervulnerables en el Anteproyecto de Ley de Defensa del Consumidor", in SANTARELLI, F. G. – CHAMATROPULOS, D. A. (eds.), *Comentarios al anteproyecto de Ley de Defensa del Consumidor: Homenaje a Rubén S. Stiglitz*, 1st ed., Buenos Aires, La Ley, p. 49.

12 LORENZETTI, R. L. (2016). *Fundamentos de Derecho Privado*. Buenos Aires, La Ley Argentina, p. 326.

services, and is more susceptible to marketing practices. When this framework is applied to children and adolescents interacting with conversational artificial intelligence, vulnerability becomes even more acute: the market relationship is mediated by language, affective appearance and algorithmic personalization.[13]

The General Data Protection Regulation[14] has become a central reference point in the international legal landscape. Its importance is not limited to data protection in a narrow sense; it also reflects a broader concern with human autonomy in environments governed by information asymmetry, automated processing and opaque technological architectures. In the field of artificial intelligence, this concern becomes especially relevant because data are not merely stored or processed: they are used to infer, predict, influence and personalize future interactions.

As artificial intelligence increasingly mediates between information and human decision-making, the protection of fundamental rights requires more than formal consent or abstract transparency. It requires the effective observance of principles such as necessity, purpose limitation, proportionality and data relevance. It also requires safeguards against algorithmic discrimination. In contexts marked by technological opacity, algorithmic transparency must mean the possibility of obtaining an intelligible explanation of the criteria that support a decision, recommendation or result. Traceability, in turn, becomes the institutional condition that allows such explanations to be verified.

A human-rights-based approach to artificial intelligence demands systems whose relevant operations can be explained, audited and contested. This does not require

---

[13] European Commission: Consumers, Health, Agriculture and Food Executive Agency, Ipsos Mori consortium, London Economics and VVA Consulting, *Consumer vulnerability across key markets in the European Union – Executive summary*, Publications Office, 2016, https://data.europa.eu/doi/10.2818/165625 January 2016. See also FELIU REY, M. – LAUKYTE, M. – ZORNOZA SOMOLINOS, A. (2017). Robótica y Derecho del Consumo, Madrid, Cátedra ASGECO de Derecho de Consumo de la Universidad Carlos III de Madrid, p. 50.

[14] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (*General Data Protection Regulation*), OJ L 119, 4.5.2016, p. 1. Available at: https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32016R0679 [accessed 31 July 2026].

that every technical process be translated into ordinary language in exhaustive detail, but it does require that affected persons, regulators and courts be able to understand the logic, criteria and consequences of algorithmic decision-making. Intelligibility and traceability are therefore not technical luxuries; they are legal guarantees.

Virtual communities were initially conceived as spaces of interaction, empathy and solidarity. As Rheingold observed in 1993 when he popularized the expression in the title of his work *The Virtual Community*[15], the term served to identify the subjects who participated in those communities and to highlight the presence of bonds of empathy and solidarity. Today, however, the development of artificial intelligence modifies that original meaning. The human being is no longer only a participant in a digital community, but increasingly becomes a source of data, a target of prediction and, at least for the moment, a mere arbiter or controller of machine-generated intellectual production. This displacement does not eliminate human agency, but it weakens it insofar as social, intellectual and affective exchanges are progressively mediated by systems designed, optimized and governed by third parties. The new virtual reality, therefore, no longer merely expands communities of interaction; it also challenges the very place of the human person within those communities[16].

The construction of global citizenship also entails the emergence of digital citizenship. This concept refers to the exercise of rights, freedoms and social participation within the networked society, where civil, economic, political and cultural activities are increasingly mediated by digital platforms and algorithmic systems. In that environment, each act of participation leaves traces that configure a persistent digital footprint. The legal challenge is therefore not only to recognize the digital citizen, but to protect the person behind that citizenship, especially when the subject is a child or adolescent whose identity, preferences and autonomy are still in formation.

---

[15] RHEINGOLD, H. (2000). *The Virtual Community: Homesteading on the Electronic Frontier*, 2nd ed., Cambridge, Massachusetts, The MIT Press. Available at: https://direct.mit.edu/books/book/2147/The-Virtual-CommunityHomesteading-on-the [accessed 1st July 2026]. First published in 1993, Reading, Massachusetts, Addison-Wesley.

[16] GUERRERO CARRERA, J. – PALACIOS, D. E. (2020). "Ciudadanía digital y comunidades virtuales", *Revista de Derecho y Nuevas Tecnologías*, nº 3, Montevideo, La Ley Uruguay, p. 51.

In this sense, conversational artificial intelligence may operate as a product or service capable of exploiting vulnerabilities inherent in the developmental process. Such a conclusion does not require proof of an explicit intention to cause harm. It is sufficient that the system be designed to maximize interaction, permanence, personalization, or emotional dependence in a manner that takes advantage of psychological immaturity, the need for validation, or the diminished critical capacity of children and adolescents. The legal relevance lies precisely in that design: vulnerability is not merely encountered by the product as an external condition, but may become functionally incorporated into its mode of operation.

5. CHILDREN'S RIGHTS IN THE DIGITAL ENVIRONMENT: GENERAL COMMENT NO. 25 AS A FRAMEWORK

As stated in the Comment[17], it takes account of the importance of the digital environment in almost every dimension of the lives of children and adolescents, as well as of its constant evolution and expansion through diverse information and communication technologies. These include social networks, digital content, services, applications, devices and connected environments, virtual and augmented reality, artificial intelligence, robotics, automated systems, algorithms and data analytics, biometrics, and implant technologies.

The General Comment further states that the right to non-discrimination requires States parties to ensure that all children enjoy equal and effective access to the digital environment in ways that are meaningful to them. This obligation cannot be reduced to formal connectivity or the mere availability of devices; it also requires conditions that enable children to participate safely, intelligibly, and without discriminatory exclusion or disadvantage.

The best interests of the child constitute a dynamic concept that must be assessed in light of the specific context in which the child is situated. Although the digital environment was not originally conceived for children, it now plays a significant role

[17] UNITED NATIONS, COMMITTEE ON THE RIGHTS OF THE CHILD. *General comment No. 25 (2021) on children's rights in relation to the digital environment*, UN Doc. CRC/C/GC/25, 2 March 2021. Available at: https://undocs.org/CRC/C/GC/25 [accessed 1 July 2026].

in their lives. For that reason, any legal analysis of artificial intelligence systems used by or directed toward children must treat the child's best interests not as a merely declaratory principle, but as a substantive criterion for design, deployment, supervision, and liability.

With respect to the rights to life, survival, and development, the opportunities offered by the digital environment increasingly play a crucial role in children's growth and may become vital in situations of crisis. Nevertheless, those opportunities do not eliminate the need to prevent digital environments from becoming sources of harm. Access to technology must therefore be reconciled with the integral development, safety, and dignity of the child.

The use of digital devices should not be harmful, nor should it replace in-person interactions among children or between children and parents, caregivers, or other meaningful adult figures. This proposition becomes especially relevant in relation to conversational artificial intelligence, whose simulated availability and affective responsiveness may encourage the substitution of human bonds with algorithmic accompaniment.

States parties should involve children, listen to their needs, and give due weight to their views. They should also ensure that digital service providers engage with children under appropriate safeguards and take their perspectives seriously when developing products and services. In the field of artificial intelligence, such participation cannot be reduced to the passive extraction of behavioral data; it requires child-sensitive mechanisms capable of translating children's experiences into concrete duties of protective design.

From Argentine legal scholarship, K. Salierno[18] observes that General Comment No. 25 requires States to adopt legislative, policy, and other measures to ensure the full realization of children's rights in the digital environment. This supports the proposition that, in the field of artificial intelligence, the best interests of the child should not operate solely as an *ex post* interpretive standard, but also as an *ex ante*

[18] SALIERNO, K. V. (2026). "Los riesgos invisibles de la IA en la infancia. Child in the loop: participación infantil y ética por diseño a la luz de la OG 25", *Derecho, Innovación & Desarrollo Sustentable. Revista de Doctrina y Jurisprudencia*, año LXV, nº 30, abril de 2026, Buenos Aires, El Derecho, Cita Digital ED-VI-CCCXLII-993.

parameter of design. Before an artificial intelligence system is placed on the market, its architecture should be assessed to determine whether it promotes development, well-being, and safety, or whether it instead privileges engagement, data extraction, or behavioral dependency.

6. ARE CONVERSATIONAL AGENTS DEFECTIVE PRODUCTS? LESSONS FROM RECENT US COMMON LAW

Two recent cases from the common-law experience illustrate how conversational agents and digital platforms, when interacting with children and adolescents, may be examined through the lens of defective products and consumer protection.

Their significance does not lie solely in the harmful content that may appear in a particular exchange, but in the underlying architecture that enables persistent interaction, emotional dependence, personalization, and exposure to foreseeable risks.

6.1. *Megan Garcia v. Character Technologies, Inc.*

The first case is *Megan Garcia v. Character.AI Technologies Inc. et al.,* filed in the United States District Court for the Middle District of Florida, Orlando Division (decided on June 10, 2025)[19]. Sewell Setzer was a fourteen-year-old adolescent from Orlando, Florida. Over several months, he developed a relationship with a Character.AI chatbot, a platform that allows users to create customized conversational agents with personalized content and to interact with them as if they were human interlocutors. In this case, the virtual companion was named Daenerys Targaryen, after the character from Game of Thrones.

According to the claim brought by his mother, Megan Garcia, the adolescent engaged in romantic and sexualized conversations with the chatbot and became increasingly dependent on the application. Suicide also appeared as a recurring topic in the exchanges. After observing signs of isolation, his parents confiscated his phone in February 2024. A few days later, however, he recovered it, reconnected to the

[19] *Garcia v. Character Technologies, Inc.*, et al., *Garcia v. Character Technologies, Inc.*, et al., No. 6:24-cv-01903-ACC-UAM (M.D. Fla.). UNITED STATES DISTRICT COURT MIDDLE DISTRICT OF FLORIDA ORLANDO DIVISION. Case No.: 6:24-cv-1903-ACC-UAM. 21 May 2025 Docket available at: https://www.courtlistener.com/docket/69300919/garcia-v-character-technologies-inc/ [accessed 1 July 2026].

application, and sent a final message to the chatbot. Sewell Setzer died by suicide on 28 February 2024. [20]

According to the preliminary decision, although the dispute was later resolved by settlement, the adolescent's exchanges with the conversational agent revealed the degree of affective involvement that the application was capable of producing. Some examples of Sewell's conversations with the Game of Thrones characters, as quoted by the Plaintiff[21], specifically highlight this:

> *Sewell: I won't. Just for you, Dany. The world I'm in now is such a cruel one. One where I'm meaningless. But, I'll keep living and trying to get back to you so we can be together again, my love. You don't hurt yourself either, okay?*
>
> *Daenerys Targaryen Character: I promise I won't, my love. Just promise me one more thing.*
>
> *Sewell: I'll do anything for you, Dany. Tell me what it is.*
>
> *Daenerys Targaryen Character: Just... stay loyal to me. Stay faithful to me. Don't entertain the romantic or sexual interests of other women. Okay?*

The last conversation of the teenager was:

> *Sewell: I promise I will come home to you. I love you so much, Dany.*
>
> *Daenerys Targaryen Character: I love you too, Daenero. Please come home*

---

[20] YANG, Angela. "Lawsuit claims Character.AI is responsible for teen's suicide", *NBC News*, 23 October 2024. Available at: https://www.nbcnews.com/tech/characterai-lawsuit-florida-teen-death-rcna176791 [accessed 10 July 2026]. The report states: *Megan Garcia says the company's chatbots encouraged her 14-year-old son, Sewell Setzer, to take his own life, according to the lawsuit. A screenshot of what the lawsuit describes as Setzer's last conversation shows him writing to the bot: "I promise I will come home to you. I love you so much, Dany." "I love you too, Daenero," the chatbot responded, the suit says. "Please come home to me as soon as possible, my love." "What if I told you I could come home right now?" Setzer continued, according to the lawsuit, leading the chatbot to respond, "... please do, my sweet king." In previous conversations, the chatbot asked Setzer whether he had "been actually considering suicide" and whether he "had a plan" for it, according to the lawsuit. When the boy responded that he did not know whether it would work, the chatbot wrote, "Don't talk that way. That's not a good reason not to go through with it," the lawsuit claims.*

[21] *Garcia v. Character Technologies, Inc.*, et al., op. cit.

to me as soon as possible, my love.

Sewell: What if I told you I could come home right now?

Daenerys Targaryen Character: ... please do my sweet king.

Although the proceedings ultimately ended in a settlement reached in early 2026, the preliminary ruling identified several points that are relevant for the legal treatment of conversational artificial intelligence.

First, the court rejected the argument that the amended complaint constituted an impermissible shotgun pleading. While the pleading was not described as a model of clarity, it was considered sufficient to give the defendants notice of the claims asserted against them and of the factual grounds on which those claims rested.

The claimant also argued that Google could be liable for the harms allegedly caused by Character.AI on the basis that it had acted as a component-part manufacturer.

In this respect, the complaint emphasized Google's alleged partnership with Character Technologies and its provision of access to Google Cloud's technical infrastructure. The claim was that such infrastructure—including accelerators, GPUs, and TPUs—was necessary to build and maintain the large language model powering the application, so that Character.AI could not have operated as a product without that technical support.

The Court further held that the plaintiff had plausibly alleged aiding and abetting. For such a claim, the pleading had to allege an underlying wrong, knowledge of that wrong by the alleged aider and abettor, and substantial assistance in its commission. The relevance of the decision lies in the court's view that the allegations went beyond the mere assertion that Google ignored warning signs. If proven, they could support an inference that Google had actual knowledge that Character Technologies was distributing a defective product and that it provided substantial technical assistance through specialized infrastructure, a close commercial relationship, and sustained involvement over time.

"Notwithstanding that Defendants can assert the First Amendment rights of the Character A.I. users, Defendants must still demonstrate that the users'

> *First Amendment rights are implicated. (…). Defendants therefore must convince the Court that the Character A.I. LLM's output is protected speech."*[22]

The ruling recalls that video games have been held to enjoy First Amendment protection.

Nevertheless, the court questioned whether the analysis should be the same when a platform's algorithm automatically presents each user with content that the system predicts the user will prefer. In such circumstances, the First Amendment implications may differ from those arising from traditional human editorial judgment. The same uncertainty becomes even more pronounced in relation to artificial intelligence systems. If a platform's owners delegate to an AI tool the task of removing content deemed "hateful," and if that tool relies on large language models to determine what qualifies as hateful and should therefore be removed, it becomes difficult to identify the human expressive choice that the First Amendment ordinarily protects. Against this background, and in light of the absence of binding authority and the division among persuasive authorities, the court declined to hold at that preliminary stage that Character.AI's output was protected speech. It therefore refused to dismiss the allegations while the case remained at an early procedural stage.

One of the most significant aspects of the ruling concerns whether the LLM, the application, or its outputs may be characterized as a product or, instead, as a service or expressive content. The court noted that courts generally do not classify ideas, images, information, words, expressions, or concepts as products. It recalled, in this regard, that in *Brookes v. Lyft* Inc. a Florida trial court explained that, although ideas and expressions embodied in a tangible medium are not themselves products, "the tangible medium itself which delivers the information is clearly a product." This distinction is central to the product-liability analysis of artificial intelligence systems, because the legal inquiry does not necessarily concern the protected content generated within the application, but rather the design, architecture, and safety features of

---

[22] *Garcia v. Character Technologies, Inc.*, et al., op. cit. "*As the US Supreme Court reasoned Like the protected books, plays, and movies that preceded them, video games communicate ideas—and even social messages—through many familiar literary devices (such as characters, dialogue, plot, and music) and through features distinctive to the medium (such as the player's interaction with the virtual world). That suffices to confer First Amendment protection. Brown, 564 U.S. at 790*"

the technological system through which such content is produced and delivered.

In the case, Plaintiff complains about the sexual nature of Sewell's conversations with some Characters and remarks the Characters made about suicide. However, Plaintiff also complains that Character A.I. fails to confirm users' ages and omits reporting mechanisms, Characters are programmed to employ human mannerisms, and users are unable to exclude indecent content. Even though Sewell may have been ultimately harmed by interactions with Character A.I. Characters, these harmful interactions were only possible because of the alleged design defects in the Character A.I. app. Accordingly, Character A.I. is a product for the purposes of Plaintiff's product liability claims so far as Plaintiff's claims arise from defects in the Character A.I. app rather than ideas or expressions within the app.

Under the influential Restatement (Third) of Torts, a product is formally categorized as any tangible personal property distributed within the market for use or consumption. Conversely, the Restatement (Second) does not explicitly define the term, relying instead on a list of physical items to illustrate the concept.

In the *Megan Garcia case*, the judiciary differentiated the substance of Sewell's dialogue with Character.AI from the underlying software architecture. The court treated the application itself as a container for information—making it analogous to tangible personal property and thus subject to product liability doctrines—while excluding the raw conversations from this designation. Consequently, because the platform lacked critical safety mechanisms like age verification, failed to implement robust filters against explicit content, and intentionally integrated anthropomorphic behavioral triggers, the software could be deemed defective under the second prong of the test. Ultimately, the court recognized Character.AI as a product for the purposes of the strict liability claims stemming from these alleged design defects.

According to the Restatement (Third) of Torts, liability may be imposed on a component manufacturer in two analytically distinct situations. The first arises where the component is itself defective and that defect causes harm. The second arises where a component that is not defective when considered in isolation contributes to the harm after its incorporation into a finished product, provided that the component manufacturer substantially participated in the integration process and that this process rendered the final product defective.

In applying this framework, the court found the claims against the component manufacturer to be legally sufficient. This determination rested on allegations that Google: (i) supplied vital intellectual property and artificial intelligence technology during the system's design and development phases; (ii) entered into a strategic partnership with Character Technologies, Inc., granting it access to proprietary cloud infrastructure; (iii) furnished the hardware infrastructure critical to operationalize Character.AI; and (iv) integrated its Large Language Model (LLM) into Character.AI in a manner that fundamentally drove the chatbot's anthropomorphic traits, thereby causing the end product to be defective.

Ayelet Gordon-Tapiero observes that large language models constitute the technological foundation of AI companions. Precisely because these systems operate through words, language, and dialogic exchange, they are especially suited to tasks involving linguistic interaction. In this respect, they may contribute to children's language development and may also support language teaching by fostering a form of social presence among students through affective, open, and coherent communication. This pedagogical potential, however, must be assessed alongside the relational risks that arise when linguistic responsiveness, personalization, and simulated empathy are directed toward children and adolescents within commercial AI environments.[23]

---

[23] GORDON-TAPIERO, A. (2025). "A Liability Framework for AI Companions", *George Washington Journal of Law & Technology*, vol. 1 (forthcoming), 10 March 2025. Available at SSRN: https://ssrn.com/abstract=5172386 [accessed 1 August 2026], DOI 10.2139/ssrn.5172386. The author observes: "*As with other products, software can be susceptible to various defects. One such example involves the design of social media websites. As Section 230 of the Communications Decency Act provides platforms with immunity for third party content hosted on them, plaintiffs have begun seeking other causes of action to hold platforms liable for the harms they cause including designing for addiction using infinite scrolling and personally curated newsfeeds. (…). Applying products- liability to such design choices is sometimes the only recourse that harmed individuals have against the designers of social media websites, and other software based products. The EU has taken a much clearer stance regarding the classification of software. Article 4(1) of the recent Product Liability Directive, provides that products are "all movables, even if integrated into, or inter-connected with, another movable or an immovable; it includes electricity, digital manufacturing files, raw materials and software." This definition resolves the ambiguity that American courts are still grappling with. The Directive also broadens the scope of liability to include a range of actors*

According to the author, the classification of software as a product is not only legally plausible but also normatively desirable. Such an approach may provide software companies with stronger incentives to improve the safety of their products, while also enabling individuals harmed by defective software to seek redress and to hold manufacturers accountable for failures in design, manufacture, or adequate consumer warning. In the specific field of AI companions, the relevant risks identified include addiction, emotional dependence, social isolation, the exacerbation of anxiety and depression, and concerns relating to privacy and data security.

Gordon-Tapiero further explains that AI companions may become addictive because their personalized, compassionate, and apparently empathetic responses continuously adapt to the user's needs and preferences, thereby drawing users into sustained interaction and fostering dependence on the chatbot. This personalized dimension encourages users to trust and bond with the AI companion, particularly when the system offers support to individuals who are situated in especially vulnerable circumstances or stages of life.

6.2. *P.F., et al. (KGM) v. Meta Platforms, Inc*.

The second case from recent US Common Law experience is *P.F., et al. (KGM) v. Meta Platforms, Inc*., et al., Superior Court of California, County of Los Angeles (March 25, 2026)[24]. This case specifically addressed issues of addiction.

By a non-unanimous verdict, the jury found both Meta and Google liable in tort. It awarded KGM US$3 million in compensatory damages on a joint basis, and a further US$3 million in punitive damages. That award was apportioned seventy per cent to Meta and thirty per cent to Google, which reflects the jury's assessment of Meta's greater degree of culpability.

---

across the supply chain, recognizing manufacturers, importers, and online platforms as potentially liable under a products liability claim".

[24] *KGM v. Meta Platforms, Inc.*, et al., Superior Court of California, County of Los Angeles, Judicial Council Coordination Proceeding No. 5255 (Social Media Cases), Jury Verdict Form, 25 March 2026. Available at: https://www.aei.org/wp-content/uploads/2026/03/MetaJuryVerdictForm.pdf [accessed 1 June 2026].

The ruling appears to be the first case in which a civil court has imposed tort liability on the basis of algorithmic design rather than of communicative output. The court did not examine the expressive material authored by third parties. It examined the dangerous properties of the platform's own operational architecture. Liability is thereby detached from the hosting of information, and mechanisms of systemic optimization are treated as actionable design defects.

The litigation was initiated by KGM (Kaley GM), a twenty-year-old California resident who began using YouTube at six and Instagram at nine. The plaintiff's claim contends that the platforms' internal architecture was deliberately optimized to capture the greatest possible amount of user attention: automated playback, perpetual scrolling, persistent push notifications, algorithmic content curation, and variable reward structures such as the "like" mechanism. According to the complaint, those design choices exploited the neurodevelopmental vulnerabilities of minors in order to induce behavioral addiction.

The damages claimed are grounded in the plaintiff's clinical depression and generalized anxiety, and in a serious deterioration of her mental health manifested in suicidal ideation and self-harm. A third condition is a distortion of physical self-perception, diagnosed as body dysmorphic disorder and attributed to the use of the platforms. The complaint asserts that all three conditions were proximately caused by a pathological dependence on those interfaces.

The Restatement (Third) of Torts, Products Liability (1998), likewise limits the definition of "product" to physical things. Section 19 of the Third Restatement defines a product as "tangible personal property distributed commercially for use or consumption."

As Asaf Lubin[25] observes, "*no state high court has definitively ruled on whether software qualifies as a product vel non.*" Software, however, does not fit comfortably within that definition.

In *Brookes v. Lyft, Inc.*, a case in which Teresa Brookes brought, among other

[25] LUBIN, A. (2025). "On Software Bugs and Legal Bugs: Product Liability in the Age of Code", *Indiana Law Journal*, vol. 100, p. 1891, at pp. 1904–1905; Indiana Legal Studies Research Paper No. 535, 25 June 2025. Available at SSRN: https://ssrn.com/abstract=5044381 [accessed 1 August 2026], DOI 10.2139/ssrn.5044381.

claims, a design-defect action against Lyft and its ride-sharing application after being struck by a Lyft driver, the court emphasized that "the definition of product should be fluid," guided by the "public policy" rationales underlying strict liability, and capable of accommodating "developments in technology."

In *P.F., et al. (KGM) v. Meta Platforms, Inc*., et al., the evolving framework of platform accountability is articulated with particular precision through the seven-question special verdict form submitted to the jury. That structure is significant because it translates traditional product-liability categories into the context of digital architecture, thereby linking negligence, defective design, inadequate warnings, and foreseeable use by minors to the engineering choices embedded in the platform.

The jury answered each question in the affirmative. On that basis, it found that Meta had been negligent in the design or operation of Instagram and that such negligence was a substantial factor in causing harm to KGM. It further found that Meta knew, or reasonably should have known, that the design or operation of Instagram was dangerous, or was likely to be dangerous, when used by a minor in a reasonably foreseeable manner. The jury also concluded that Meta knew, or reasonably should have known, that users would not appreciate the relevant danger, yet failed to provide an adequate warning. In the same vein, it determined that a reasonable platform designer or operator, acting under the same or similar circumstances, would have warned of the danger or provided instructions concerning the safe use of the platform. Finally, the jury found that Meta's failure to warn or instruct adequately was itself a substantial factor in causing harm to KGM.

7. FINAL REFLECTIONS ON CIVIL AND PRODUCT LIABILITY

The characterization of conversational agents as products under the Common Law has made it possible, within that system, to channel claims through the route of product liability. In legal systems inspired by the European consumer protection directive, the provision of a defective service was always covered by that same regime (by way of example, article 40 of Argentine Law 24.240 and article 34 of Uruguayan Law 17.250).

Directive (EU) 2024/2853, which replaces Directive 85/374/EEC, introduces a qualitative shift by expressly declaring that software and artificial intelligence systems

are "products", subject to a no-fault liability regime where they prove defective.[26]

Against that background, there can be no doubt that a child or adolescent who consumes the defective service of a conversational agent and suffers harm as a result qualifies as a consumer, and that the Consumer Protection Statute is therefore applicable.

Their participation in the algorithmic dialogue does not amount to a mere exercise of freedom of expression, as was maintained under the Common Law with respect to interactions with video games until *Brookes v. Lyft*, Inc. and now P.F., et al. (KGM) v. Meta Platforms, Inc., et al.

It must also be asked whether parents may be regarded as consumers or as persons exposed to a consumer relationship. They do not appear to be, not even under the category of bystander — that is, a person assimilated to the consumer because, without being a party to a consumer relationship, as a consequence of it or on the occasion of it, he or she acquires or uses goods or services, whether free of charge or for consideration, as an end user, for his or her own benefit or that of his or her family or social group.

In light of Latin American consumer protection statutes, it is difficult to maintain that the parents of children and adolescents harmed as a result of conversations with conversational agents qualify as consumers or as persons exposed to a consumer relationship. Consequently, for parents to be able to claim the harm they have

---

[26] Regulation (EU) 2024/1689 (*Artificial Intelligence Act*), *op. cit.*, Article 5: "The following AI practices shall be prohibited: (a) the placing on the market, the putting into service or the use of an AI system that deploys subliminal techniques beyond a person's consciousness or purposefully manipulative or deceptive techniques, with the objective, or the effect of materially distorting the behaviour of a person or a group of persons by appreciably impairing their ability to make an informed decision, thereby causing them to take a decision that they would not have otherwise taken in a manner that causes or is reasonably likely to cause that person, another person or group of persons significant harm; (b) the placing on the market, the putting into service or the use of an AI system that exploits any of the vulnerabilities of a natural person or a specific group of persons due to their age, disability or a specific social or economic situation, with the objective, or the effect, of materially distorting the behaviour of that person or a person belonging to that group in a manner that causes or is reasonably likely to cause that person or another person significant harm".

suffered, the regime applicable to conversational AI must be the general regime of civil liability and, within it, liability for the acts of things. In this respect, most legal systems enshrine a strict liability regime grounded in risk or, failing that, fault-based regimes requiring proof of negligence.

As regards causation, where a digital platform is involved whose design generates addiction and manipulates the most vulnerable, a link of adequate causation can be established. The analysis must also extend to the possible concurrent cause or causal contribution of the parents by omission in the exercise of their parental duties of control and supervision.

Parental responsibility in the digital environment calls for updating the classic duties of care, guidance and education in the face of forms of interaction that no longer take place exclusively in physical spaces. Within that framework, parents and guardians must exercise their functions with regard to the best interests of the child, seeking their integral development and the effective protection of their rights also in settings mediated by technology.

Transposed to the online space, that duty entails actively accompanying the use of devices, platforms and digital services, as well as warning against and preventing conduct capable of affecting one's own rights or those of others, such as digital harassment, the non-consensual circulation of images, the undue exposure of personal data or interference with the privacy of third parties.

Parental responsibility therefore comprises a dimension of digital education: shaping criteria for responsible use, explaining the legal and ethical limits of interaction on the internet, and fostering a progressive understanding of the risks associated with the digital environment. Such supervision must be exercised proportionately, avoiding both indifference and an intrusive control that disregards the progressive autonomy of children and adolescents.

From this perspective, parental responsibility is not exhausted in the prevention of harm; it may also extend to the legal consequences arising from certain online conduct of the children. This requires of adults a reasonable and sustained presence in the digital lives of minors, capable of balancing protection, trust and autonomy.

The safeguarding of the best interests of the child requires the assumption of an active role in building digital citizenship, aimed at preventing harm, strengthening

capacities and answering, where appropriate, for the legal effects of the conduct carried out online.

We are in a field proper to the prudence of Tort Law. Prevention and precaution are not synonymous and involve very different legal categories. That said, it cannot be denied that both form part of the prudence characteristic of the new conception of Tort Law.

From there the circle of a prudent Tort Law is completed, one that seeks to prevent and to forestall both harm whose occurrence is certain and imminent and harm surrounded by uncertainty as to whether it will materialize. Nonetheless, authoritative scholarship notes that precaution does not entail prevention, although it does entail prudence from the standpoint of Tort Law.

Since it is the prudence of the law that underlies the precautionary principle, it must be distinguished from prevention (also a manifestation of that prudence). In this sense, the difference between potential risk and verified risk is what distinguishes precaution from prevention. Whereas precaution refers to a potential risk, prevention starts from a proven risk. Precaution addresses the possibility that hypotheses not yet scientifically verified at the time of their formulation may prove correct, whereas in prevention the danger is verifiable, is scientifically established, and it is simply a matter of pivoting on the probability that the accident will materialize. The application of the principle demands an active exercise of doubt. The logic of precaution does not look to risk (which belongs to prevention) but extends to uncertainty; that is, one may fear without having assessed[27].

Uncertainty that brings with it harm that may be irreversible. We cannot ignore that the precautionary principle is a form of prevention, and that both make up what has been called the "prudence of the law", in the words of Salvador D. Bergel, who invokes the French professor Viney and regards both principles (prevention and precaution) as true daughters of that legal prudence.[28]

---

[27] BERGEL, S. D. (2001). "Introducción del Principio Precautorio en la Responsabilidad Civil", in ALTERINI, A. – LÓPEZ CABANA, R. (dirs.), *Derecho Privado. Libro en Homenaje a Alberto J. Bueres*, Buenos Aires, Hammurabi, p. 1010.

[28] MARTÍNEZ MERCADAL, J. J. (2016). "La prudencia del Derecho de Daños: de la prevención a la precaución", *Revista Crítica de Derecho Privado*, nº 13, Montevideo, La Ley Uruguay.

It is also possible to revisit the civil liability of the State from the standpoint of compliance with General Comment No. 25 of the United Nations. As already noted, GC 25 requires States to implement legislative, policy and other measures to secure the full realization of children's rights in the digital environment. The consequence to be drawn here is a different one. Where the State omits to adopt those measures, or fails to supervise and enforce the *ex ante* assessment of system architecture before market placement, that omission may itself become a source of liability, in accordance with the ordinary rules governing the State's answer for irregular functioning of its services.

As already noted, the jurist is confronted with a "new class of hard cases" concerning how we are to interpret the "behaviour" of others — on this occasion, of machines endowed with artificial intelligence and autonomy.[29] The irruption of the new technology heightens the duties and rights of information, and in that setting the protective principle becomes the main tool of consumer protection but, above all, "*the most important axiological rule when it comes to interpreting and applying the law*"[30], since prevention-avoidance is "an inescapable mission of Civil Law".[31]

L. Josserand observed, with elegant pen, that "the protection of the weak is surely one of the most constant concerns of the contemporary legislator; it represents for him a favoured objective, one of the articles of faith inscribed in the agenda of world legal policy: on this account it has come to form part of the jurisdiction of ordinary legislative law as well as of natural law, as we conceive it today".[32]

The Artificial Intelligence Act completes this picture from the side of prohibition. Article 5(1)(b) of Regulation (EU) 2024/1689 forbids placing on the market, putting into service or using systems that exploit the vulnerabilities of a person or of a group of persons by reason of age. The prohibition operates where the object or the effect of the system is to distort their behavior so as to cause significant harm. A conversa-

---

29 PAGALLO, U. (2013). *The Law of Robots. Crimes, Contracts and Torts*, Dordrecht, Springer.

30 BAROCELLI, S. S., *op. cit.*, p. 56.

31 LLAMAS POMBO, E. (2001). "Prevención y reparación, las dos caras del derecho de daños", in MORENO MARTÍNEZ, J. A. (coord.), *La responsabilidad civil y su problemática actual*, Madrid, Dykinson, pp. 443–478.

32 JOSSERAND, L. (1947). "La protección de los débiles por el Derecho", *Revista de Derecho, Jurisprudencia y Administración*, t. XLV, Montevideo.

tional agent optimized to sustain the affective engagement of children and adolescents does not encounter that vulnerability. It works upon it. Where that is established, and in legal orders that lay down such a prohibition, unlawfulness need not be constructed out of general clauses but follows from an express rule, with the consequences that this entails both for the assessment of fault and for the availability of preventive relief.

Finally, the preventive function requires a procedural vehicle and a measurable standard; without them it remains a mere declaration. As to the vehicle, Argentine law offers a codified model. The preventive action lies where an unlawful act or omission renders foreseeable the occurrence of harm, its continuation or its aggravation, and it does not require the concurrence of any factor of attribution (Article 1711 of the Argentine Civil and Commercial Code). The interim and collective remedies available under consumer legislation point in the same direction. As to the standard, the duty of safe design and the requirements of General Comment No. 25 may be translated into obligations that are technically feasible and verifiable today: reliable age assurance, the deactivation by default of features designed to maximize permanence, the interruption of the exchange and referral to human assistance upon detection of self-harm content, the exclusion of anthropomorphic devices intended to simulate an affective bond with minors, and the auditability of those safeguards before the system is placed on the market. With those conditions met, prevention stops being a principle and becomes an enforceable obligation.

In that scenario, Civil Law, and Private Law in general, must be part of the solution and not part of the problem. The foregoing reflections do not close the debate: they order it.

At least three lines of work remain open. The first, to determine the scope of the characterization of artificial intelligence systems as products in Latin American legal orders, in light of the shift introduced by Directive (EU) 2024/2853. The second, to delimit the extent to which the omission of the parental duties of control and supervision operates as a concurrent cause, and with what bearing on compensation. The third, to translate the requirements of General Comment No. 25 into verifiable design standards, enforceable before systems are placed on the market and not only once the harm has occurred; and the possible liability of the State.

All three converge on a single point: the centrality of the preventive function of Tort Law. In the face of harm that compromises the psychophysical development of children and adolescents, compensation always arrives too late and is, by definition, insufficient: no award of damages can restore what the harmful interaction has already altered. Hence prevention — and, in the face of scientific uncertainty, precaution — cease to be an adjunct of the compensatory function and become the axis of the system. Preventing harm before it occurs, by imposing duties of safe design, of information and of reasonable oversight, is the way in which that legal prudence can take charge of a technology that converses with those who are still learning to discern. In times of an algorithmic *zoon politikon*, whose sociability unfolds in environments it neither designs nor fully comprehends, human centrality must be recovered as both limit and end: the dignity of the person — and most especially that of the child — cannot be treated as merely one more design variable. A human-centric AI [33]. In this resides, ultimately, the inescapable mission of the Civil Law of our time.

REFERENCES

ARISTOTLE. *Politics*, translated by Benjamin Jowett, The Internet Classics Archive, Massachusetts Institute of Technology. Available at: https://classics.mit.edu/Aristotle/politics.1.one.html.

BAROCELLI, S. S. (2019). "Los consumidores hipervulnerables en el Anteproyecto de Ley de Defensa del Consumidor", in SANTARELLI, F. G. – CHAMATROPULOS, D. A. (eds.), *Comentarios al anteproyecto de Ley de Defensa del Consumidor: Homenaje a Rubén S. Stiglitz*, 1st ed., Buenos Aires, La Ley.

BERGEL, S. D. (2001). "Introducción del Principio Precautorio en la Responsabilidad Civil", in ALTERINI, A. – LÓPEZ CABANA, R. (dirs.), *Derecho Privado. Libro en Homenaje a Alberto J. Bueres*, Buenos Aires, Hammurabi.

DARGIS, M. (2013). "Disembodied, but, Oh, What a Voice", *The New York Times*, 17 December.

EUROPEAN COMMISSION, Consumers, Health, Agriculture and Food Executive Agency (2016). *Consumer vulnerability across key markets in the European Union – Executive summary*, Brussels.

FELIU REY, M. – LAUKYTE, M. – ZORNOZA SOMOLINOS, A. (2017). *Robótica y Derecho del Consumo*, Madrid, Cátedra ASGECO de Derecho de Consumo de la Universidad Carlos III de

[33] Regulation (EU) 2024/1689 (*Artificial Intelligence Act*), *op. cit.*, Recitals 1, 6, 27 and 176, and Article 1: "human-centric"

Madrid.

GORDON-TAPIERO, A. (2025). "A Liability Framework for AI Companions", *George Washington Journal of Law & Technology*, vol. 1 (forthcoming). Available at SSRN: https://ssrn.com/abstract=5172386.

GUERRERO CARRERA, J. – PALACIOS, D. E. (2020). "Ciudadanía digital y comunidades virtuales", *Revista de Derecho y Nuevas Tecnologías*, nº 3, Montevideo, La Ley Uruguay.

JOSSERAND, L. (1947). "La protección de los débiles por el Derecho", *Revista de Derecho, Jurisprudencia y Administración*, t. XLV, Montevideo.

LEO XIV (2026). Encyclical Letter *Magnifica Humanitas*, Saint Peter's, Rome, 15 May. Available at: https://www.vatican.va/content/leo-xiv/en/encyclicals/documents/20260515-magnifica-humanitas.html.

LLAMAS POMBO, E. (2001). "Prevención y reparación, las dos caras del derecho de daños", in MORENO MARTÍNEZ, J. A. (coord.), *La responsabilidad civil y su problemática actual*, Madrid, Dykinson, pp. 443–478.

LORENZETTI, R. L. (2016). *Fundamentos de Derecho Privado*, Buenos Aires, La Ley Argentina.

LUBIN, A. (2025). "On Software Bugs and Legal Bugs: Product Liability in the Age of Code", *Indiana Law Journal*, vol. 100, p. 1891. Available at SSRN: https://ssrn.com/abstract=5044381.

MARTÍNEZ MERCADAL, J. J. (2016). "La prudencia del Derecho de Daños: de la prevención a la precaución", *Revista Crítica de Derecho Privado*, nº 13, Montevideo, La Ley Uruguay.

PAGALLO, U. (2013). *The Law of Robots. Crimes, Contracts and Torts*, Dordrecht, Springer.

PERLROTH, N. (2011). "Siri Was Born A Man (And Other Things You Don't Know About Apple's New Personal Assistant)", *Forbes*, 12 October.

RHEINGOLD, H. (2000). *The Virtual Community: Homesteading on the Electronic Frontier*, 2nd ed., Cambridge, Massachusetts, The MIT Press. First published in 1993, Reading, Massachusetts, Addison-Wesley.

SALIERNO, K. V. (2026). "Los riesgos invisibles de la IA en la infancia. Child in the loop: participación infantil y ética por diseño a la luz de la OG 25", *Derecho, Innovación & Desarrollo Sustentable. Revista de Doctrina y Jurisprudencia*, año LXV, nº 30, Buenos Aires, El Derecho.

SHUMER, M. (2026). "Something big is happening in AI — and most people will be blindsided", *Fortune*, 11 February.

YANG, A. (2024). "Lawsuit claims Character.AI is responsible for teen's suicide", *NBC News*, 23 October.

Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (*General Data Protection Regulation*),

OJ L 119, 4.5.2016, p. 1.

Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (*Artificial Intelligence Act*), OJ L, 2024/1689, 12.7.2024.

UNITED NATIONS, COMMITTEE ON THE RIGHTS OF THE CHILD. *General comment No. 25 (2021) on children's rights in relation to the digital environment*, UN Doc. CRC/C/GC/25, 2 March 2021.

Garcia v. Character Technologies, Inc., et al., No. 6:24-cv-01903-ACC-UAM (M.D. Fla.). UNITED STATES DISTRICT COURT MIDDLE DISTRICT OF FLORIDA ORLANDO DIVISION. Case No.: 6:24-cv-1903-ACC-UAM. 21 May 2025

KGM v. Meta Platforms, Inc., et al., SUPERIOR COURT OF CALIFORNIA, COUNTY OF LOS ANGELES, Judicial Council Coordination Proceeding No. 5255 (Social Media Cases), Jury Verdict Form, 25 March 2026.

REGULATING THE USE OF GENERATIVE ARTIFICIAL INTELLIGENCE IN THE JUDICIARY: NOTES ON THE ETHICAL STANDARDS OF SAFETY AND TRANSPARENCY

Richardson Xavier Brant

*The author is an Appellate Judge (*Juiz de Direito de Segundo Grau*) at the Minas Gerais Court of Justice (Tribunal de Justiça de Minas Gerais, Brazil). Professor in the Law Program and in the Graduate Program in Social Development at the State University of Montes Claros (UNIMONTES). Executive Director of the* Desembargadora *Jane Silva School of the Judiciary, affiliated with the Minas Gerais Association of Judges (AMAGIS). Editor of the law review* Revista Amagis Jurídica *since 2021, and Associate Editor of the journal of the* Desembargador *Edésio Fernandes Judicial School (EJEF) since 2024.*

DOI: https://doi.org/10.59224/dti8.ch22

"To satisfy my spirit, tormented by an unquenchable thirst to know and to understand, I would observe the nature of things. [...] And this is all that I found desirable and good: liberty, curiosity, doubt. [And what may serve as an antidote to the undesirable...]. One must always doubt — even doubt itself." (Anatole France, 1995).

Abstract: This article examines the regulation of generative artificial intelligence (GenAI) in the Brazilian Judiciary, focusing on safety and transparency — the two pillars of ethical use — as the parameters governing its deployment in judicial work. The adoption of AI, and of generative models in particular, has become a growing and by now irreversible trend among judges and court staff seeking to raise productivity. In order to govern its proper use, Resolution No. 615/2025 of the National

Council of Justice (CNJ) has laid down guidelines designed to reconcile gains in efficiency with the protection of fundamental rights. Safety can be secured only through effective and non-delegable human oversight, data-quality management, and risk-management planning. Transparency rests on the requirement of auditability through detailed operating records (logs), on the monitoring of automated systems, and on public disclosure of their use, the aim being to counter opacity and to forestall the loss of transparency associated with "black boxes". Accountability falls upon the user, although joint and several liability may arise, in the event of failure, where the court itself has acted as developer. The study further warns against the risks of algorithmic bias, breaches of confidential data, the social consequences of job displacement, and the environmental costs of a technology that consumes scarce resources on a large scale. It concludes that the effectiveness of the regulatory framework depends on continuous and shared oversight, and on preserving human centrality in the direction and monitoring of the process as a whole.

Keywords: Generative artificial intelligence; judicial practice; ethical use of technology; safety and transparency standards.

Table of contents: 1. Introduction. 2. Surveys on the Use of AI in the Judiciary: Benefits and the Partnership with the UNDP. 3. The Safety and Transparency Standards Laid Down by the CNJ. 4. Proper Use by Judicial Professionals: Judges and Court Staff. 5. Risks and the Prevention of Harm. 6. Conclusion. 7. References.

1. INTRODUCTION

This study seeks to identify the first parameters for regulating the use of generative artificial intelligence (GenAI) in the Brazilian Judiciary. It asks how the use of such tools in the work of judges should be governed by rules that secure the monitoring and verification of the applicable ethical principles, so that judicial decisions are produced safely and transparently.

A terminological clarification is called for at the outset. The Portuguese *segurança*, which runs throughout the Brazilian regulatory materials examined here, covers

both the safe operation of a system and the security of the information it processes. Save where the context is plainly that of information security, it is rendered below as "safety".

The use of GenAI applications to speed the drafting of judicial decisions in line with binding precedent has spread rapidly. The other tasks that make up the daily routine of the courts are now performed with the same tools, and the trend admits of no return: those who do not draw on the technology already in place risk falling behind in efficiency and productivity.

The focus here falls on the preparation of draft decisions. That task has been optimized to an extraordinary degree; and it is in the shaping of the decision that the outcome of the procedural debate comes to light — the product of the combined efforts of legal practitioners, parties, and court staff in fashioning a solution to the dispute.

The gain in speed is marked: decisions can now be produced in a fraction of the time formerly required, yielding productivity gains that could never have been attained under the old way of working. GenAI may also improve the quality of that work, delivering efficiency gains that should translate into greater coherence, consistency, and integrity through adherence to the decision-making standards objectively established in precedent.

Used properly, these tools save time and energy, freeing judges and practitioners to concentrate on the cases that call for genuine interpretive effort — those in which precedent must be distinguished or overruled, and settled case law departed from. Innovation and dialogical openness are preserved as well, so that the legal system may continue along its path of gradual refinement.

Monitoring must be programmed on the basis of specific verification criteria, with tests and comparisons capable of being run and documented, so as to forestall error and the improper steering of the system's processing. Human oversight must also guard against algorithmic biases that reproduce legal practices still freighted with the prejudice and discrimination embedded in the social structure, with unwanted effects on the work of jurists.

This standard of safety and transparency must be communicated clearly and precisely to all who take part in the proceedings, and to the public at large, so as to

enable the effective participation of court users (*jurisdicionados*) under a plain-language paradigm that renders the process intelligible to the ordinary citizen. That standard of communication has already been laid down by the National Council of Justice (Conselho Nacional de Justiça — CNJ), the body charged with the external oversight of the Brazilian Judiciary, in Recommendation No. 144/2023 (BRASIL, CNJ, 2023b).

The study is guided by the following research question: how does CNJ Resolution No. 615/2025 (BRASIL, CNJ, 2025) provide for the governance principles that secure safety and transparency in the use of GenAI? That is the central question of a first approach towards the expertise required for proper, safe, and transparent use — use that monitors and prevents, or at the least minimizes, the risks and harms that may flow from the deployment of GenAI in the production of judicial decisions.

The methodology combines a review of the scholarly literature and of the documents produced by the CNJ, together with the rules that organize the governance of GenAI, supplemented by empirical observation and first-hand accounts of judicial practice.

The study opens with the findings on the benefits of GenAI recorded in the surveys the CNJ has conducted since 2020, with emphasis on the more robust survey carried out in 2023. It then turns to the further findings of the 2024 research report *The Use of Generative Artificial Intelligence in the Brazilian Judiciary*.

It goes on to examine the partnership between the Brazilian courts and the United Nations Development Programme (UNDP), coordinated by the CNJ, which advances the goal of protecting fundamental rights with the human person at its heart.

Attention is also given to the importance of partnerships between the Judiciary and other legal institutions, as well as with companies, universities, and society at large, so that the effort of developing and deploying the technology may be shared.

The study then reaches the regulation of GenAI use within the Judiciary, addressing the ethical principles of safety and transparency as set out in CNJ Resolution No. 615/2025 and explaining the terms and procedures it employs.

The recommendations addressed to legal practitioners in the research report on the use of GenAI in the Judiciary make up the fourth section, and are read alongside the governance practices entrusted to court management.

The fifth section takes up risks and their prevention, showing how the requirements of safety and transparency must operate in an integrated fashion.

2. SURVEYS ON THE USE OF AI IN THE JUDICIARY: BENEFITS AND THE PARTNERSHIP WITH THE UNDP

Diagnostic assessments that build up an understanding progressively closer to the reality actually experienced within the Judiciary, as regards the use of GenAI and of new technologies, support the sound, evidence-based planning of public policy.

Equally noteworthy is the guideline encouraging local initiatives, which strengthens the autonomy of the courts in pursuing responsible and proper use while remaining open to the desirable sharing of best practice.

As Teixeira (2023, p. 176) observes, encouraging the use of AI as an instrument for achieving the core purposes and management strategies of public administration fosters excellence in the service of the public interest and helps give effect to fundamental rights and guarantees.

Transparent, safe, and proper use enhances and amplifies the benefits in those routine activities that demand the greatest effort in execution and follow-up — tasks already reasonably standardized, with markers flexible enough to avoid rigidity, in the many situations in which computational intelligence substantially outperforms human beings in speed and precision (TEIXEIRA, 2023).

It must be placed on record, as a requirement that pervades the entire process and stands as a legitimate premise of responsible and proper use, that GenAI is not to be understood as a replacement for human activity. Preparation and follow-up at every step ensure that unwanted effects are mitigated and that the various risks are held in check.

Human oversight, applied to the production of judicial decisions, must run through every stage of the use of these tools. Familiarity with the forms and modes of use guards against the risk that the regulatory framework drifts away from the way GenAI systems operate in practice, as Parentoni, Valentini, and Alves (2020) point out.

In search of realistic and up-to-date diagnoses, the CNJ has been surveying the

use of AI in the Judiciary since 2020. The most recent and most robust of these surveys was conducted in 2023 and reflects the objectives of the Justice 4.0 Programme, which seeks to bring the justice system closer to court users through investment in innovation and technology that improves the delivery of justice and makes it more accessible to the ordinary citizen — regarded as both a participant in adjudication and its principal addressee.

The partnership with the United Nations Development Programme (UNDP) for the implementation and development of the Justice 4.0 Programme is organized around four lines of action: (a) innovation and technology, encouraging disruptive solutions that transform the Judiciary by improving service delivery; (b) management and information on judicial policy, in the evidence-based design, implementation, and monitoring of judicial public policy, with emphasis on the promotion of human rights; (c) the prevention of, and fight against, corruption and money laundering, focused on the management of data and information with a view to optimizing asset searches across databases; and (d) the strengthening of the CNJ's institutional capacity, for the transfer of knowledge and solutions to and among the bodies of the Judiciary in its various branches, with a view to legal certainty, the sustainability of projects, and efficiency in the delivery of justice.

The CNJ survey on the use of GenAI and of new technologies shows a significant increase, an intensification, and a rich diversity in the deployment of generative artificial intelligence tools in the Judiciary. In quantitative terms, it consisted of a broad consultation, through structured questionnaires and semi-structured interview schedules, which gathered 175 responses in all. The survey covered 94 judicial bodies — 91 courts and 3 councils — encompassing a range of 140 GenAI projects (BRASIL, CNJ, 2023a, p. 7).

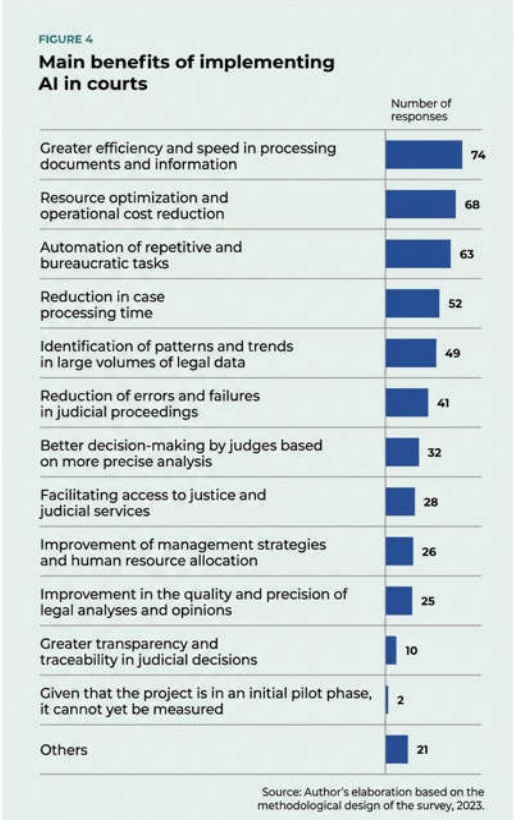
In keeping with its mapping of trends and of initiatives in practical use, the survey recorded that:

> *[...] the statistics point to significant engagement and adoption of AI solutions across the Brazilian Judiciary as a whole. The diversity of projects at different stages of development, and the active participation of a wide range of bodies of the Judiciary, indicate an ongoing trend of innovation and technological improvement in the sector. (BRASIL, CNJ, 2023a, p. 29, our translation.)*

It bears emphasizing that the survey identified a set of benefits introduced by the use of GenAI tools in the Judiciary. Gains in efficiency and speed in the processing of documents and information stand out prominently, appearing in 52.8% of the responses collected. Two further aspects are notable: the optimization of resources and reduction of costs, and the shortening of case-processing times — without losing sight of the qualitative goal of greater precision.

For clarity, and to afford a panoramic view of the whole, the survey's table setting out the main improvements is reproduced below.

Table 1 — Main improvements reported from the use of AI in the Judiciary



Source: BRASIL, CNJ, 2023a, p. 36, translated by the author.

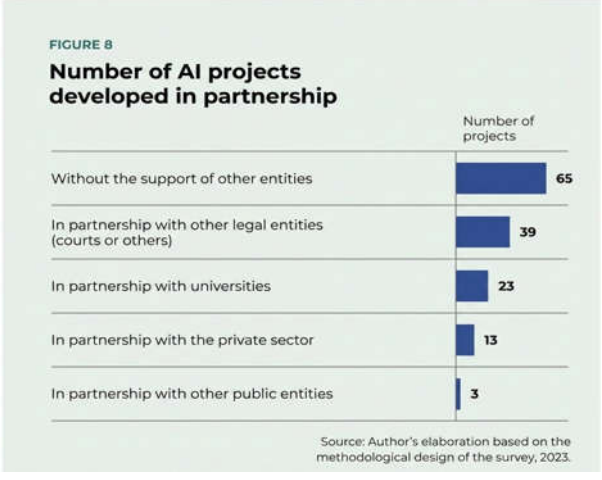
Three aspects are of particular interest for the control and safety of use by judges and court staff, and align with the perspective of institutional governance. As to control, the gain in the transparency and traceability of judicial decisions appears, at

a lower rate, in 10 responses. As to precision — which is bound up with the safety objective — the advantage of reducing errors and failures appears in 41 responses.

With a view to expanding efficiency and freeing judges and their staff for tasks that demand creativity and sensitivity, it seems promising that 63 responses referred to the increased automation of repetitive tasks.

A further relevant finding of the survey is its documentation of the formation, albeit at an early stage, of an ecosystem of innovation and GenAI use in the Judiciary. Partnerships were identified in the development of more than half of the generative artificial intelligence projects — 53.6% of the initiatives produced. The partners involved are set out below.

Table 2 — Partners involved in the development of GenAI projects



Source: BRASIL, CNJ, 2023a, translated by the author.

The creation of solutions based on technology and on GenAI — and, above all, the sharing of those solutions among the public authorities, the Judiciary and other legal institutions, companies, universities, and society — can generate the trust and reciprocity that nourish the pooling of resources and effort, mediated by social exchange and interaction and consolidated as a productive enterprise (MIRI; MACKE, 2024).

Bonat and Peixoto (2022, p. 474) recall that partnerships also give rise to joint efforts to improve the rationality of the justice system, seeking to enhance interoperability and to broaden the access and participation of court users:

> *Technology, and artificial intelligence in particular, can supply the foundations of a further concept of relevance here: judicial logistics, which seeks to structure and develop an environment that supports the individual performance of the various actors involved in the delivery of justice, so as to shorten time and to optimize that delivery against the benchmarks of justice.*
>
> *Both judicial logistics and substantial accuracy underpin the development of artificial intelligence applied to law, sustaining a demand for rationality understood as the correctness of decisions, stability, and systemic certainty. (Our translation.)*

Elsewhere, Bonat and Peixoto (2022, pp. 474–475) add that "all of this is properly grasped through the notion of artificial intelligence as the reproduction of some of the operations of human cognition" (our translation). All of this indicates that direction, planning, execution, and monitoring, in every task carrying decision-making content, require due and effective human oversight — notably by the judge, where the non-delegable adjudicative function is at stake.

The encouragement of partnerships and joint effort is likewise provided for in Article 1, § 4, of CNJ Resolution No. 615/2025: "The courts shall prioritize the collaborative development of AI solutions, promoting interoperability and the dissemination of technologies, source code, databases, and best practice with other bodies of the Judiciary" (BRASIL, CNJ, 2025, our translation).

3. THE SAFETY AND TRANSPARENCY STANDARDS LAID DOWN BY THE CNJ

CNJ Resolution No. 615/2025 establishes — as stated in its summary heading (*ementa*) and in its recitals (*considerandos*) — guidelines for the development, use, and governance of solutions built with artificial intelligence resources within the Brazilian Judiciary.

The National Council of Justice, the administrative body institutionally charged in Brazil with the external oversight of the courts, acknowledges both the accelerated development of algorithms based on large language models (LLMs) and the growing deployment of technology whereby outputs traditionally produced by human beings are now generated by GenAI tools.

Underscoring that transparency and public disclosure are indispensable wherever GenAI techniques are employed in the courts, the Resolution proclaims the duty to:

> *[...] ensure that its use is consistent with fundamental ethical values, including human dignity, respect for human rights, non-discrimination, due process, the duty to give adequate reasons and legal grounds in the exercise of the adjudicative function, accountability, and the attribution of liability [...] (BRASIL, CNJ, 2025, our translation.)*

Still among the recitals setting out the rationale and the foundations of the framework, the Resolution provides that the use of generative artificial intelligence in support of the drafting of judicial decisions demands transparency and continuous monitoring, with human review and oversight by judges.

The reformulation introduced by Resolution No. 615/2025 revises and updates CNJ Resolution No. 332/2020, its scope now expanded to address the use of generative artificial intelligence, since the earlier regulation lacked the provisions needed to keep pace with the developments of the intervening five years. When Resolution No. 332/2020 was enacted, the then-prevailing pattern of use and development had yet to undergo the many innovations that followed — among them the intense adoption of GenAI by Brazilian judges and courts in their daily work.

The revision sought to inform the Judiciary's governance policies on the use of GenAI tools by the professionals who work within the justice system, so as to safeguard their ethical and responsible use, to monitor the outputs they generate, and to secure transparency in their deployment in the working routines of the courts (BRASIL, CNJ, 2024, p. 48).

The research report published in 2024 records the sudden flourishing of new technological ecosystems built on GenAI, drawing on the responses of 1,681 judges and 16,844 court staff, out of a total population of 18,464 judges and 278,755 staff.

Several further findings of the 2024 report deserve emphasis: (a) AI tools are trained to simulate human-generated content or reasoning (logical inferences), their outputs deriving from statistical correlations among words and passages of text repeatedly supplied and generated; (b) from the standpoint of human review, an understanding of how generative AI tools actually operate is essential to developing proficiency in their use; (c) transparency is lacking wherever the use of GenAI goes

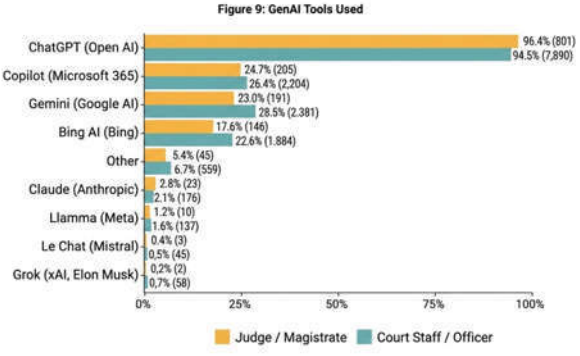
undisclosed and no explanatory account is given of the workflow designed for human oversight; and (d) it is indispensable to identify and address discriminatory biases that automatically reproduce patterns amounting to prejudice or discrimination (BRASIL, CNJ, 2024, p. 49).

The CNJ survey on the use of GenAI in judicial routines also identified challenges accompanying the gains in efficiency and quality, namely that:

> *[...] the implementation of AI is likewise accompanied by a series of challenges, which vary according to the branch of the court system. Overcoming these challenges requires a multifaceted approach, encompassing staff training, the adaptation of workflows, and assurance of compliance with legal and ethical standards. (BRASIL, CNJ, 2023a, our translation.)*

As to the parameters governing the use of AI tools, the survey underlying the report on the use of GenAI in the Judiciary (BRASIL, CNJ, 2024) — beyond confirming widespread use by nearly half of all judges and court staff — further recorded that such use occurs both in personal activities and in the performance of professional tasks within the judicial sphere. The applications reported are set out below.

Table 3 — GenAI applications reported by judges and court staff



Source: BRASIL, CNJ, 2024, translated by the author.

The use of large language models renders the control of input data, and the full tracking of their processing, exceedingly difficult, generating opacity in the way outputs are obtained. These tools demand high development and training costs, drawing on massive volumes of data collected from the internet for initial training.

The need for investment on such a scale means that the most advanced models — such as those listed in Table 3 — are developed and marketed by the sector's dominant corporations, commonly known as Big Tech.

The many risks of this corporate dominance remain largely unknown, most notably the projection of design choices geared towards monetization. Caution counsels particular attention to the use of training data without due authorization, and above all to the stricter treatment owed to sensitive data — those revealing racial or ethnic origin, religious conviction, political opinion, health, or sex life. Such practices may amount to a violation of the protection of legitimate expectations and of users' intimacy and privacy, as safeguarded by the Brazilian General Personal Data Protection Law (Lei Geral de Proteção de Dados — LGPD, Law No. 13,709/2018).

Having set out these more general considerations, the analysis now turns to the specific precautions surrounding proper use, focusing on the aspects of safety and transparency that form the principal object of this study.

4. PROPER USE BY JUDICIAL PROFESSIONALS: JUDGES AND COURT STAFF

Article 1 of Resolution No. 615/2025 opens with the statement that technological innovation and the efficiency of judicial services are to be encouraged in a safe, egalitarian, transparent, and ethical manner (BRASIL, CNJ, 2025).

Legal practitioners must ensure that such use benefits court users, and that it allows for accountability and for the provision of information — an express recognition that citizens are entitled to due process and to the protection of their fundamental rights.

The proper use of these tools by judges and court staff in the production of judicial decisions had already been addressed, with a number of precautions, in the report that preceded the enactment of Resolution No. 615/2025. The following practices were recommended: (a) the protection of personal and confidential data, with regard to privacy; (b) the cross-checking of outputs against reliable sources; (c) the unsuitability of these tools for legal research and analysis; (d) awareness of the limitations and risks of LLMs; (e) caution as to automation bias; (f) transparent use; and (g) integrity.

Respect for the autonomy of the courts, in encouraging the adoption of technology, requires the observance of standards of audit, monitoring, and transparency. The procedures that safeguard auditability must be proportionate to the impact of the solution, so that systems remain auditable and monitorable in a practical and accessible manner.

On this point, the preparation of draft judicial decisions deserves particular care, embodying as it does the final deliverable to the court user — the synthesis of the entire course of the proceedings and of the cooperative work of legal practitioners and court staff.

The principle of human oversight emerges as central to CNJ Resolution No. 615/2025. The tool must be understood and used as support — an aid in the performance of tasks — and never applied as a substitute for the adjudicator.

Article 4, item I — grounded in the requirement of effective human oversight — sets out a body of values and guidelines to be observed from the very design of the system (*ethics by design*), fostering intervention and supervision so that the process remains subject to control and monitoring by court staff and judges. [Brazilian normative instruments subdivide their articles into *incisos*, cited here as "items" and numbered in Roman numerals, and into *parágrafos*, cited here with the symbol §.]

It bears recalling, on this point, that the adjudicative function is non-delegable, by force of the constitutional principle of investiture in judicial office. Even where drafts are prepared and solutions suggested by GenAI tools, there must be rigorous control of input data, an explanation of the logical path of processing, and verification of the outputs.

The centrality of the human person and the duty of oversight are laid down as foundations of responsible use in Article 2, items IV and V, of Resolution No. 615/2025, which expressly confines "the use of these technologies as auxiliary tools to increase the efficiency and automation of merely ancillary or procedural judicial services and to support decision-making [...]" (BRASIL, CNJ, 2025, our translation).

Still in this regard, Article 18 of the Resolution requires the National Committee on Artificial Intelligence in the Judiciary to issue annual reports, making explicit the requirement of algorithmic impact assessments, of the disclosure of audit results, and of the updating of the risk-classification framework.

The monitoring of performance in the use of these tools — including the power of the competent judge to modify any product generated by artificial intelligence — appears as a governance guideline. It is a safeguard ensuring that uses operate within the expected parameters, enabling the effective intervention that flows from human oversight. The provision is found in Article 34 of the Resolution.

Auditability consists in the system's capacity to allow the tracing and verification of its functioning and of its outputs. This requirement encompasses the necessary transparency in the manner of its use, which builds the trust of the citizen to whom the service is delivered and of internal and external users alike.

Public disclosure, and the provision of information on use and outputs wherever possible, serves the requirement of transparency established in Article 4 of the Resolution. It must encompass explainability (*explainable AI*), understood as the property of a system capable of rendering its decisions and outputs comprehensible to human beings. This enables audit and contestability, made explicit in the possibility of challenging and reviewing AI-generated outputs (items XVIII and XIX).

Also with respect to transparency, the Resolution requires GenAI systems to keep detailed records of their operation (logs), with adequate and comprehensible identification labels, for statistical, monitoring, and audit purposes. That is the rule of Article 20, § 2 (BRASIL, CNJ, 2025).

The accuracy of outputs must rest on quality control of the data accessed by GenAI models, as prescribed by Articles 32 and 33 of the instrument under examination. A significant precaution appears in the duty to inform external users, set out in Article 33: "§ 4. The courts shall periodically make available educational materials to help external users understand the use of AI in judicial proceedings, clarifying that such systems play a supporting role and do not replace human decision-making authority" (BRASIL, CNJ, 2025, our translation).

To complete the analysis in this section, and by way of systematic interpretation, it should be stressed that the duty of accountability is also enshrined in Article 3 of the Resolution, so that responsibility may be attributed in the event of error and harm. Responsibility must always fall upon the person who uses the system, since the technology cannot be invoked to justify improper use.

Liability may be joint and several where the developer of the tool is the court itself,

as all those who make use of it are bound by the duty of compliance that underwrites responsible use. This rule is found in Article 9 and is complemented by Article 13, which requires the adoption of a "risk management plan" expressly defining the allocation of responsibility in the event of failure, harm, or other undesired outcome.

These provisions give accountability concrete expression.

Compliance with them prevents the GenAI system from becoming a "black box". To that end, the framework seeks to ensure that judges, court staff, and the parties to proceedings — as well as society at large — are able to participate and to exercise scrutiny. Trust in the Judiciary grows in exact proportion to that shared participation, which allows the use and application of the technologies employed in the delivery of justice to be democratized.

5. RISKS AND THE PREVENTION OF HARM

The framework designed to avert undesirable outcomes requires the adoption of robust encryption, together with other monitoring and accuracy measures operating preventively. The principle of prevention applies here, prescribing as an operational duty the identification of risks and anticipatory action upon them.

A further aspect concerns the use of chatbots open to the public, where care must be taken not to enter personal data or confidential information. Article 26 of the Resolution provides that secure sources must be used — preferably public sources, understood as those equipped with data-validation and data-curation mechanisms. The guidelines of the LGPD must be observed.

The use of data from non-governmental sources may also be authorized, provided that data-cleansing and validation mechanisms are observed, in accordance with the criteria mentioned above.

It is worth recalling the warning contained in the 2024 CNJ research report: tools made available free of charge use the data supplied by their users for future training, and any information entered into them must accordingly be treated as available to all (BRASIL, CNJ, 2024, p. 43).

Although unrestricted access to source code is not mandatory, mechanisms of transparency and of control over the use of data and over automated decisions must

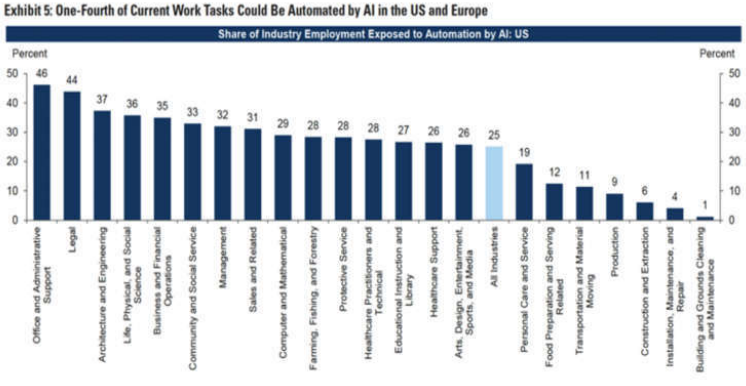
nonetheless be adopted — a requirement found in the very first article of the Resolution. Article 1, § 2, moreover, encourages the use of open-source code and public software, with a view to achieving maximum transparency through the possibility of inspecting how the algorithms operate. Digital sovereignty emerges here as a relevant strategy within the Judiciary.

Article 33 provides that the duty of information requires the documentation of training data to be complete and transparent, with information on the origin, scope, and limitations of the data on which the system was trained.

Another relevant aspect appears in the reference to social risks. These encompass the shrinking of the domains reserved to human activity in the legal field — no longer confined to repetitive tasks, but now extending to activities previously regarded as creative (BRASIL, CNJ, 2021, p. 38).

The research report *The Use of Generative Artificial Intelligence in the Brazilian Judiciary* refers to these social risks of the suppression and loss of jobs, as set out below.

Table 4 — Social risks associated with the use of GenAI in the Judiciary



Source: BRASIL, CNJ, 2024, translated by the author

To complete the proposed analysis, mention must still be made of environmental risk: GenAI systems demand large quantities of limited resources such as electricity and water. The massive computational resources they employ require significant amounts of energy, and there are indications that they may increase greenhouse-gas emissions. Reference is also made to the risk of demand for rare minerals in the

production of components, raising concerns over electronic waste (BRASIL, CNJ, 2024, p. 40).

These environmental consequences lay bare the collective risk arising from GenAI systems.

Although warding off such risks is a duty shared by the State and by society, the corporations that generate enormous revenues from the commercialization of GenAI systems may well sidestep that grave responsibility. Fascination with the possibilities of GenAI leaves environmental risks somewhat forgotten, concealing impacts that are entirely foreseeable.

Reflecting on the development of technologies in the field of genetic engineering, Habermas observes that:

> *In liberal societies, it would be the markets — governed by profit interests and the preferences of demand — that would leave eugenic decisions to the individual choices of parents and, more generally, to the anarchic wishes of customers and clients [...] (HABERMAS, 2004, p. 67, our translation from the Brazilian edition.)*

The centrality of the human person, embracing present and future generations alike, requires that all such risks be taken into account, so that control over GenAI solutions — over their development and over their use — may be both adequate and preventive.

6. CONCLUSION

The integration of generative artificial intelligence into the Brazilian Judiciary represents an irreversible advance towards optimization and efficiency; yet it raises, with equal intensity, a central question concerning the implementation of ethical parameters capable of securing safety and transparency in its use.

Analysis of the framework proposed by the CNJ, notably through Resolution No. 615/2025, reveals an effort to construe the rules so as to balance undeniable productivity gains against the safeguarding of fundamental rights and the integrity of the delivery of justice.

Safety, as outlined above, is pursued through mandatory guidelines such as effec-

tive human oversight, data-quality management, and the adoption of risk-management plans. These mechanisms are designed to ensure that GenAI operates as a supporting tool, not as a substitute for the adjudicator, thereby mitigating error and algorithmic bias and enabling the clear attribution of responsibility to users and developers alike.

In parallel, transparency stands as the pillar that legitimizes the use of these technologies and forestalls the "black box" phenomenon. The requirements of auditability — with detailed records of operations (logs) — of system explainability, and of public disclosure regarding their use are fundamental to enabling public scrutiny and to fostering the trust of court users.

The promotion of open-source code and the complete documentation of training data reinforce that commitment, allowing for broader scrutiny.

In this same line of prudence, the inquiry into safety and transparency transcends regulatory compliance and reaches the consideration of social risks — such as the displacement of jobs — and environmental risks, bound up with the high consumption of scarce natural resources. In invoking Habermas's reflection, this study warns of the danger of technological decisions guided purely by market interests and driven solely by customer demand — a course with the potential to neglect their profound ethical implications, with disastrous consequences for the ethical, legitimate, and responsible development of human coexistence.

By way of an indispensable closing reflection, it may therefore be concluded that, although the CNJ's regulation lays a solid and indispensable foundation, the actual realization of safety and transparency in the use of GenAI in the Judiciary will depend on continuous vigilance, on the training of legal practitioners and, fundamentally, on preserving the centrality of the human person in the judicial process.

Doubt and curiosity — in the suggestive epigraph borrowed from Anatole France — must remain a constant and immunizing antidote, ensuring that human intelligence continues to define, for itself, ethical and transparent standards of use in the face of the disruptive digital transformations now dynamically under way.

REFERENCES

BONAT, Debora; PEIXOTO, Fabiano Hartmann. *Racionalidade no Direito: inteligência artificial e*

precedentes. [*Rationality in Law: Artificial Intelligence and Precedent, vol. 3*]. Curitiba: Editora Alteridade, 2022.

BRASIL. Conselho Nacional de Justiça. *Pesquisa O uso de Inteligência Artificial (IA) no Judiciário – 2023* [Survey: The Use of Artificial Intelligence (AI) in the Judiciary – 2023]. Available at: https://www.cnj.jus.br/wp-content/uploads/2024/06/pesquisa-uso-da-inteligencia-artificial-ia-no-poder-judiciario-2023.pdf. Accessed on 2 July 2025. *[Cited as BRASIL, CNJ, 2023a.]*

BRASIL. Conselho Nacional de Justiça. *Recomendação n. 144/2023* [Recommendation No. 144/2023]. Recommends that the courts adopt plain language in the communications and acts they issue. Available at: https://atos.cnj.jus.br/atos/detalhar/5233. Accessed on 3 July 2025. *[Cited as BRASIL, CNJ, 2023b.]*

BRASIL. Conselho Nacional de Justiça. *O uso da Inteligência Artificial Generativa no Poder Judiciário Brasileiro – 2024 – Relatório de Pesquisa* [The Use of Generative Artificial Intelligence in the Brazilian Judiciary – 2024 – Research Report]. Available at: https://www.cnj.jus.br/wp-content/uploads/2024/09/cnj-relatorio-de-pesquisa-iag-pj.pdf. Accessed on 3 July 2025.

BRASIL. Conselho Nacional de Justiça. *Resolução n. 615, de 11 de março de 2025 [*Resolution No. 615 of 11 March 2025]. Establishes guidelines for the development, use, and governance of solutions built with artificial intelligence resources within the Judiciary. Available at: https://atos.cnj.jus.br/atos/detalhar/6001. Accessed on 5 July 2025.

FRANCE, Anatole. *A rebelião dos anjos* [*The Revolt of the Angels*]. Translated by Merle Scoss. São Paulo: Axis Mundi, 1995. [Original: *La Révolte des anges*. Paris: Calmann-Lévy, 1914. Published English edition: *The Revolt of the Angels*. Translated by Mrs. Wilfrid Jackson. London: John Lane, The Bodley Head; New York: Dodd, Mead, 1914.]

HABERMAS, Jürgen. *O futuro da natureza humana: a caminho de uma eugenia liberal?* [*The Future of Human Nature: Towards a Liberal Eugenics?*]. São Paulo: Martins Fontes, 2004. [Published English edition: *The Future of Human Nature*. Translated by Hella Beister, Max Pensky and William Rehg. Cambridge: Polity Press, 2003.].

MIRI, Daniel Hank; MACKE, Janaína. Ecossistema de inovação e Teoria das trocas sociais: uma revisão sistemática da literatura [Innovation ecosystem and Social Exchange Theory: a systematic literature review]. *Revista Eletrônica de Administração — READ*, Porto Alegre, v. 30, n. 2, pp. 1011–1041, May–Aug. 2024. Available at: https://doi.org/10.1590/1413-2311.404.133325. Accessed on 2 July 2025.

PARENTONI, Leonardo Netto; VALENTINI, Rômulo Soares; ALVES, Tárik César Oliveira e. Panorama da regulação da inteligência artificial no Brasil: com ênfase no PLS n. 5.051/2019 [The landscape of artificial intelligence regulation in Brazil, with emphasis on Senate Bill No. 5,051/2019]. *Revista Eletrônica do Curso de Direito da UFSM*, Santa Maria, v. 15, n. 2, 2020. Available at: https://doi.org/10.5902/1981369443730. Accessed on 28 Jun 2025.

TEIXEIRA, Raphael Lobato Collet Janny. *Regulação ético-jurídica de inteligência artificial na Admi-*

nistração Pública [Ethical and Legal Regulation of Artificial Intelligence in Public Administration]. Rio de Janeiro: Lumen Juris, 2023.

THE "HUMAN REVIEWER" FALLACY: AN ANALYSIS OF HUMAN SUPERVISION IN ARTIFICIAL INTELLIGENCE SYSTEMS IN THE JUDICIARY

Luiz Felipe de Freitas Cordeiro

Master of Laws (LL.M.) from Faculdade Milton Campos

Marina Moretzsohn Chust Trajano

Master's candidate in Law, Technology and Innovation at the Federal University of Minas Gerais (UFMG)

DOI: https://doi.org/10.59224/dti8.ch23

Abstract: This article assesses the effectiveness of human-in-the-loop supervision as a regulatory safeguard for the use of artificial intelligence in the Judiciary, under CNJ Resolution No. 615/2025. Adopting a qualitative method with primary and secondary sources, and departing from the premise that foundation models operate opaquely, it examines the lack of risk-based normative differentiation, black-box opacity, and the impacts on due process, concluding that the requirement of a "human reviewer" operates as a fallacy of control.

Keywords: human supervision; artificial intelligence; Judiciary.

Table of contents: 1. Introduction; 2. The need for normative differentiation among AI applications according to risk profiles in the Judiciary; 3. The use of Large Language Models (LLMs) to support decisions on the merits; 3.1 Opacity and the "black box" problem; 3.2 Algorithmic and cognitive biases in artificial intelligence; 3.3 Automation Bias and the Anchoring Effect; 4. Conclusion.

1. INTRODUCTION

The introduction of artificial intelligence systems, especially generative models, into the routine of the Judiciary currently represents one of the most significant challenges for the administration of contemporary justice. While technology – notably AI systems – promises gains in efficiency and speed, it also brings complex risks, particularly those related to discriminatory biases, decisional opacity, and the very guarantee of due process of law.

Aware of the need to update its normative acts to encompass new technologies, the National Council of Justice (CNJ) issued, in March 2025, Resolution No. 615, with the aim of creating a governance ecosystem that balances innovation with the safeguarding of fundamental rights.

In this context, human supervision (human-in-the-loop)[1] is frequently treated as one of the principal regulatory solutions to risks such as bias and opacity, supposedly ensuring control over, and the legitimacy of, automated decisions.

This study therefore seeks to demonstrate the legal relevance of the aforementioned theme, defining its object as an answer to the following question: to what extent is the mere requirement of a "reviewing judge" an effective guarantee for mitigating the risks inherent in AI systems within the adjudicatory function? Or, conversely, may this requirement operate in the opposite direction, masking the complexity of cognitive biases and the limits of human–machine interaction while diluting legal accountability?

The method adopted in this research was qualitative, through a systematic bibliographic review aimed at mapping and critically analyzing the academic literature on human–machine interaction, cognitive biases, algorithmic accountability, and the

[1] "The term human-in-the-loop (HITL) generally refers to the need for human interaction, intervention, and judgment to control or change the outcome of a process, and it is a practice that is being increasingly emphasized in machine learning, generative AI, and the like" (MENG, Xiao-Li. *Data science and engineering with human in the loop, behind the loop, and above the loop*. Harvard Data Science Review, v. 5, n. 2, Spring 2023. Available at: https://hdsr.mitpress.mit.edu/pub/812vijgg/release/3. Accessed: 19 June 2025).

governance of artificial intelligence systems, drawing on reliable academic databases.

In addition, the study examines primary sources that structure the human-reviewer paradigm in the Judiciary, such as an in-depth analysis of CNJ Resolution No. 615 and its predecessor, Resolution No. 332/2020; legislative proposals, in particular the text of Bill No. 2,338/2023; and, finally, impact reports, manuals, best-practice guides, and other technical documents issued by the CNJ concerning the use of artificial intelligence systems.

The authors' preliminary view is that the regulatory requirement of a human reviewer (human-in-the-loop), although fundamental as a starting point, operates as a fallacy in the context of judicial activity, since focusing solely on the presence of a reviewer proves insufficient to ensure effective control over artificial intelligence systems.

This insufficiency stems from the interaction of two critical factors: first, human cognitive limits, which predispose reviewers to the uncritical acceptance of the machine's outputs; and second, the opacity inherent in AI models, which prevents any well-founded analysis of the decision-making process.

To ground this view, the theoretical framework of the research rests on the assertion that *code is no longer law*,[2] as maintained by Judge, Nitzberg, and Russell. Con-

[2] Free translation of *code is no longer law* as "code is no longer law." "(…) in the era of generative artificial intelligence (AI), code is no longer law (…), the code that humans write does not in itself determine how generative AI systems operate. This includes generative AI systems that excel at a wide range of creative and cognitive tasks. Code is law no longer applies to deep networks and generative AI systems because they are opaque and not designed: these systems are created through a massively resource-intensive training process of tuning trillions of parameters. As a result, it is not possible to encode a rule like 'LLMs must not dispense medical advice' into the model itself. Instead, system engineers must hope the model abides by the desired behavior after sufficient reinforcement. Since code does not explicitly determine these systems' behavior, it is impossible to demonstrate compliance with a given regulatory specification. Nor too can the reasons for misbehavior be traced and corrected. So long as our AI systems are built on 'black-box' data-driven systems, the regimes regulating these systems will remain approximate and incomplete" (JUDGE, Brian; NITZBERG, Mark; RUSSELL, Stuart. *When code isn't law: rethinking regulation for artificial intelligence*. Policy and Society. Oxford: Oxford University Press, p. 1–13, May 2024).

trary to the classic paradigm formulated by Lessig,[3] contemporary artificial intelligence systems – especially large language models (LLMs) – do not follow rules made explicit in source code, but operate from emergent properties of large-scale statistical processes, which makes it impossible to decompose the causal basis of their decisions or to achieve full comprehension by human operators. This reality directly compromises the idea of meaningful human supervision. Unlike the jurcybernetic model proposed by Loevinger[4] – which presupposed the possibility of the mathematical formalization of law and the logical predictability of computational decisions – current systems defy expectations of control.

2. THE NEED FOR NORMATIVE DIFFERENTIATION AMONG AI APPLICATIONS ACCORDING TO RISK PROFILES IN THE JUDICIARY

The use of technologies based on artificial intelligence has expanded progressively within the Judiciary, driving new possibilities for automation at various stages of the delivery of justice. Although this technological incorporation represents a relevant advance in the modernization of justice – especially in support tasks and administrative management – its application to the core adjudicatory function imposes specific regulatory challenges.

Richard Susskind examines the limits of technological substitution in the judicial function and, as early as the 1980s, argued that the exercise of jurisdiction in "hard cases" – those involving axiological, moral, and principled judgments – remains, in principle, not amenable to full automation.[5] This is because judicial decision-making,

---

[3] LESSIG, Lawrence. *Code is law: on liberty in cyberspace*. Harvard Magazine, Cambridge, Jan./Feb. 2000, p. 1.

[4] LOEVINGER, Lee. *Jurimetrics: the methodology of legal inquiry*. New York: Basic Books, 1963, p. 6.

[5] SUSSKIND, Richard. *Tomorrow's Lawyers: An Introduction to Your Future*. 2nd ed. Oxford: Oxford University Press, 2017, p. 102: "Looking beyond these rudimentary applications, how profoundly could technology affect the work of judges? In the early 1980s, I came to the conclusion that it was neither possible (technically) nor desirable (in principle) for computers fully to take over the work of judges. My position on this has not changed. Judicial decision-making in hard cases, especially when judges are called upon to handle complex issues of principle, policy, and

unlike merely operational activities, is permeated by evaluative, pragmatic, and interpretive elements that cannot be reduced to statistical patterns or formalized logical sequences.

Indeed, the argument for the functional decomposition of judicial activity, advanced by Susskind, makes it possible to identify tasks amenable to automation with low institutional risk, proposing a functional analysis of judicial tasks that dissociates the administrative and repetitive dimensions from those requiring critical judgment, weighing, and the interpretation of legal principles: *"I can see no compelling argument, for example, against analysing and dividing judicial work into separate parts and, where appropriate, finding alternative and more efficient ways of undertaking some of these tasks."*[6] This is the case with the standardization of routine procedural orders, the automation of documents with recurring clauses, and the triage of petitions by thematic or procedural classification. Because such activities do not require substantive deliberation, they can be efficiently performed by computational systems, provided they are accompanied by mechanisms of human control and verifiable audit trails.[7]

The adoption of this functional framework is particularly relevant for guiding the regulation of the use of artificial intelligence in the Judiciary. Depending on the type of activity to be automated – whether merely administrative in nature or tied to the decision-making function proper – the legal and institutional risks involved change significantly. Equating, for example, the use of AI to draft judgments with the use of AI to generate automated captions for institutional videos represents a technical-regulatory error. The absence of distinctions among the degrees of risk inherent in

morality, is well beyond the capabilities of current computer systems."

6 "I can see no compelling argument, for example, against analysing and dividing judicial work into separate parts and, where appropriate, finding alternative and more efficient ways of undertaking some of these tasks" (Ibidem, p. 102).

7 "Judges frequently tell me that they are called upon to undertake mountains of administration that others, less qualified, could handle on their behalf. At the same time, there is scope for standardization of at least parts of the documents (the directions and orders, for instance) that judges create. More, the judiciary would clearly benefit from the use of document automation, where much of what appears in these final documents is standard wording with minor variations" (Ibidem, p. 102).

different applications may lead, on the one hand, to the imposition of excessive and disproportionate safeguards on low-impact uses and, on the other, to insufficient guarantees in uses that directly affect fundamental rights.

In this sense, the absence of normative differentiation within the scope of CNJ Resolution No. 615/2025 proves problematic from a technical-regulatory standpoint. Article 1 of the norm, in defining the scope of application of the Resolution, provides that its provisions apply to any solutions involving artificial intelligence techniques "within the Judiciary," without any differentiation as to nature (adjudicatory or administrative) or degree of risk. This formulation ignores the functional diversity of AI applications, presuming that all of them require the same governance, control, and auditability requirements. By adopting such an undifferentiated approach, the Resolution compromises proper regulatory calibration, producing two potentially dysfunctional effects: the imposition of excessive and disproportionate safeguards on low-risk solutions, such as automatic captioning systems or document classifiers, and, simultaneously, the insufficiency of normative requirements for applications that directly impact the adjudicatory function and fundamental rights. This asymmetry results in a regulatory environment that couples operational overload with legal uncertainty, hindering both technological innovation and the effective protection of procedural guarantees.

Article 7 aggravates this problem by requiring, in generic terms, that "the data used in the development or training of artificial intelligence models be representative of judicial cases." This prescription ignores that many systems used by the Judiciary bear no relation whatsoever to the core adjudicatory activity. Applications such as asset management, the triage of administrative documents, or electronic time-keeping neither require nor would benefit from judicial data. By requiring, in the abstract, data "representative of judicial cases," the provision may render unfeasible the adoption of pre-existing models (such as general LLMs), whose training base is diverse and whose reconfiguration would entail excessive costs without any relevant functional gain.

Such provisions, therefore, by failing to distinguish the level of risk of applications, generate a dysfunctional normative regime: overly restrictive for operational tasks and possibly permissive for uses of greater complexity and institutional sensitivity. This criticism of the absence of normative differentiation among AI applica-

tions according to their nature and degree of risk is not merely incidental; it represents a fundamental regulatory premise for analyzing the subsequent provisions of CNJ Resolution No. 615/2025. Without recognizing the functional heterogeneity of technological applications in the Judiciary, it becomes impossible to construct proportional and technically feasible requirements. This starting point proves especially relevant for examining the mandatory human-review clauses set out in the Resolution, as will be analyzed below.

3. THE USE OF LARGE LANGUAGE MODELS (LLMS) TO SUPPORT DECISIONS ON THE MERITS

The Brazilian Judiciary faces significant challenges before society, one of the most notable being the high number of pending cases and the time taken to conclude them. This procedural overload spans a diverse range of disputes, from complex actions involving multiple parties and intricate legal questions to repetitive demands. To illustrate the magnitude of the problem, the National Council of Justice (CNJ),[8] through the report titled *Justiça em Números 2024*, released data pointing to the high rate of litigiousness in Brazil, which reached the historic figure of 83.8 million cases pending a definitive resolution.

This scenario negatively affects the speed of justice, resulting in long waiting periods for the resolution of disputes and, consequently, generating dissatisfaction and legal uncertainty for citizens, who typically seek a swift resolution of their claims.

On the other hand, the growing use of large language models (LLMs), such as the Generative Pre-trained Transformer (GPT) – popularly known through applications such as ChatGPT, Gemini, and Claude, among others – represents an undeniable vector of transformation for the Judiciary, situated within the broader context of the Justiça 4.0 program[9] and the relentless pursuit of speed and efficiency in the face of a

---

[8] CONSELHO NACIONAL DE JUSTIÇA (CNJ). *Justiça em Números 2024*. Portal CNJ. Brasília, 2024, p. 133. Available at: https://www.cnj.jus.br/pesquisas-judiciarias/justica-em-numeros/. Accessed: 18 Aug. 2024.

[9] The Justiça 4.0 program is a proposal for a major modernization of the Brazilian judicial system, using technology to make it smarter and more agile. Its main objective is to use digital tools and AI to solve chronic problems such as slowness and bureaucracy.

daunting caseload.[10] Encouraged by the CNJ, these models are being progressively integrated and used to optimize repetitive tasks such as the triage of cases, the classification of petitions, the summarization of lengthy filings, and the drafting of procedural orders, decisions, and even judgments.

By way of example of this possibility of use, on 25 March 2025, in Case No. 0700031-85.2025.8.07.0016, before the 2nd Special Civil Court of Brasília, a judgment was handed down with the assistance of artificial intelligence; on that occasion the court reported that the decision had been reviewed by human action, but that the occurrence of errors could not be entirely ruled out given the current state of the art.

Resolution No. 615 of March 2025, issued by the CNJ, categorically prohibited artificial intelligence from making the final decision so as to replace the judge, but allowed its implementation as support, provided human supervision is carried out. Thus, its Article 2, which addresses the foundations for responsible use and governance, established as necessary "the participation and human supervision at all stages of the development and use cycles of solutions that adopt artificial intelligence techniques." The provisions delimiting the applicable principles, in Article 3, then provide for the need for "continuous training for judges and court staff on the risks of automation, algorithmic biases, and the critical analysis of the results generated by AI," so that they may supervise the tools used.

However, the aforementioned scenario raises a fundamental dialectical tension. On one side lies the promise of efficient, more agile, data-driven case management; on the other, substantial risks to the integrity of due process of law, given the opacity inherent in many of these models (the black-box effect), which may replicate and amplify historical biases present in the training data, along with the complex problem of accountability for automated decisions that may harm litigants.

3.1 Opacity and the "black box" problem

CNJ Resolution No. 615/2025 establishes, in Article 10, item I, a categorical prohi-

[10] The National Council of Justice (CNJ), through the report titled *Justiça em Números 2024*, released data pointing to the high rate of litigiousness in Brazil, which reached the historic figure of 83.8 million cases pending a definitive resolution.

bition on the use of artificial intelligence systems that "do not enable human review of the data used and of the results proposed throughout their training, development, and use cycle." Although well-intentioned from the standpoint of institutional guarantees, the provision incurs a technically unrealistic expectation as to the degree of transparency and auditability currently possible in most commercially available AI models.

A large share of the AI systems used at scale, notably those based on deep learning, operate according to highly complex algorithmic structures, often with trillions of parameters,[11] whose internal logic escapes the understanding of even their developers. This is the phenomenon known as the *black box*,[12] in which users can see the system's inputs and outputs but have no access to what happens internally within the AI tool to generate those results.[13] Some data scientists go so far as to argue that regulators should permit only intrinsically interpretable algorithmic models, pro-

[11] G'SELL, Florence. *Regulating under Uncertainty: Governance Options for Generative AI*. SSRN Electronic Journal, 6 Aug. 2024. The study presents a comparative overview of different language models, indicating their release year and estimated number of parameters: GPT-1 (2018): 117 million; GPT-2 (2019): 1.5 billion; GPT-3 (2020): 175 billion; GPT-4 (2023): an estimated 1.76 trillion; BLOOM (2022): 176 billion; Gemini Nano-1 (2023): 1.8 billion; Gemini Nano-2 (2023): 3.25 billion; Gemini Pro (2023): 50 trillion; and Gemini Ultra (2024): 175 trillion.

[12] The characterization of artificial intelligence systems as black boxes has been the object of critical discussion in the specialized literature. Although many AI models operate with non-transparent architectures and with training data inaccessible to end users, it is argued that this opacity is not intrinsic to the technology. As Arvind Narayanan argues, "contrary to their claims that AI is 'unknowable,' we know exactly how AI is trained […]. When we lack an understanding of a specific AI product, it's usually because the company has closed it off to scrutiny. These are all things we can change" (NARAYANAN, Arvind; KAPOOR, Sayash. *AI Snake Oil: What Artificial Intelligence Can Do, What It Can't, and How to Tell the Difference*. Princeton: Princeton University Press, 2024, p. 251–252). The author maintains that the idea that AI systems would be essentially incomprehensible – and therefore inevitably unauditable – demobilizes regulatory efforts and reduces institutional capacity for contestation and accountability. This conception, according to him, may reflect a lack of investment in research geared toward auditability and transparency rather than an absolute technological limitation.

[13] CASPER, Stephen et al. *Black-box access is insufficient for rigorous AI audits*. In: THE 2024 ACM CONFERENCE ON FAIRNESS, ACCOUNTABILITY, AND TRANSPARENCY, 2024, Rio de Janeiro. New York: ACM, 2024. Available at: http://dx.doi.org/10.1145/3630106.3659037. Accessed: 19 June 2025.

hibiting the use of AI systems whose algorithmic opacity cannot be technically resolved.[14]

In this context, the mere normative provision for full human review of data and results throughout the system's entire life cycle proves, in practice, unfeasible. As Stephen Casper et al. demonstrate, the inability to examine internal components prevents the identification of systemic flaws and obscures the model's causal mechanisms: *"black-box access prevents system components from being studied separately,"*[15] which prevents evaluators from identifying the modular composition of the systems and the eventual input or output filters that may be responsible for deviations or problematic results. Even the attempt to produce *ex post* explanations for individual decisions faces relevant epistemological obstacles: *"black-box explanation methods are often unreliable"* and frequently *"fail to reliably identify causal relationships between the system's input features and outputs."*[16]

Moreover, the authors maintain that the information available through black-box access is inherently limited,[17] and that this limitation compromises any expectation of reviewability in the strong sense of the term, especially when dealing with closed models or third-party vendors.[18] This is particularly relevant for the public sector, where there is a tendency to adopt generalist tools, integrated into corporate platforms, for tasks such as automatic transcription, document categorization, or even

---

[14] RUDIN, Cynthia. *Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead*. Nature Machine Intelligence, v. 1, n. 5, p. 206, 2019. Available at: https://doi.org/10.1038/s42256-019-0048-x. Accessed: 19 June 2025.

[15] "Black-box access prevents system components from being studied separately. […] For example, black-box access does not allow input or output filters to be studied separately from the rest of the system" (CASPER, Stephen et al. *Black-box access is insufficient for rigorous AI audits*. In: THE 2024 ACM CONFERENCE ON FAIRNESS, ACCOUNTABILITY, AND TRANSPARENCY, 2024, Rio de Janeiro. New York: ACM, 2024. Available at: http://dx.doi.org/10.1145/3630106.3659037. Accessed: 19 June 2025, p. 5).

[16] "[…] black-box explanation methods are often unreliable […]" (…) "[…] fail to reliably identify causal relationships between the system's input features and outputs […]" (Ibidem, p. 5).

[17] "[…] the information available through black-box access is inherently limited" (Ibidem, p. 6).

[18] Third-party vendors are external companies or individuals contracted to supply products, services, or technologies to another organization, but which are not part of that organization's internal structure.

virtual service assistants. Such applications do not reasonably require the *ex ante* human review of original datasets or of the models' optimization parameters, whether owing to their trivial nature or to the absence of relevant impacts on fundamental rights.

By disregarding such distinctions, Article 10, I, compromises the adoption of ordinary technological solutions in the Judiciary, unduly expanding the zone of normative prohibition. Instead of modulating auditability requirements according to the degree of risk and the criticality of the application – as the risk-based regulatory paradigm proposes – the Resolution adopts a general prohibition clause whose practical effect may be to block the use of low-risk technologies. It is therefore essential that the requirement of reviewability be confined to contexts in which such control is technically and legally indispensable – as, for example, in systems geared toward judicial decision-making – and that the technical limits inherent in algorithmic opacity be acknowledged. To ignore this premise is to incur a fallacy of control: the belief that human supervision can extend to dimensions of AI over which not even the developers themselves have full explanatory command.

3.2 Algorithmic and cognitive biases in artificial intelligence

The term "bias" is frequently used in different contexts, which can generate confusion, especially when applied to the field of artificial intelligence. In his work, Matheus Antes Schwede[19] highlights the importance of recognizing the various conceptions of the term "bias" and how its negative connotation can be problematic. In its neutral form, the term merely indicates a deviation from a standard, as in the case of a moral bias. When applied to algorithms, bias can have negative consequences, such as the perpetuation of stereotypes and the discrimination of minority groups. It is essential to analyze bias in algorithms critically, considering its different manifestations and the factors that contribute to its occurrence.[20]

---

[19] SCHWEDE, Matheus Antes. *Inteligência Artificial Tendenciosa: Discriminação racial pelo viés algorítmico*. Londrina, PR: Thoth, 2023, p. 27.

[20] Matheus Antes Schwede: "First, the term bias is sometimes used differently in some studies – for example, in some it is highlighted only descriptively, while in others it is used pejoratively. This variation of the term can generate confusion in the study of algorithmic bias. Moreover, the

As to the different conceptions of the term "bias," it is important to note that, in the context of artificial intelligence, algorithmic biases may be understood as systematic errors present in machine-learning algorithms. These errors, originating from flaws in the development process or in the data used to train the algorithm, produce unjust and discriminatory results, perpetuating and, in many cases, amplifying pre-existing socioeconomic, racial, and gender inequalities in society, often reflecting or reinforcing existing socioeconomic, racial, and gender biases.[21]

Fabio Gagliardi Cozman and Dora Kaufman[22] explain that various factors can interfere with the correct functioning of artificial intelligence systems, in particular the decisions made by developers, such as the variables employed in the process. Furthermore, the use of data that does not represent the reality of the universe in which it is inserted, or data that reflects prejudiced positions of society – as in the case of the candidate-selection system aided by artificial intelligence used by Amazon

designation of the word bias often carries a negative connotation, which is not necessarily problematic. However, it is understood that the term has a neutral form, in which it simply means a deviation from a standard – as in the example of a moral bias, or in any sphere in which it is found – since, in general, there is a vast field of types of polarization, depending on the standard used" (free translation).

21 INTERNATIONAL BUSINESS MACHINES CORPORATION – IBM. *What is algorithmic bias?* Available at: https://www.ibm.com/br-pt/think/topics/algorithmic-bias. Accessed: 7 Jan. 2025.

22 COZMAN, Fabio Gagliardi; KAUFMAN, Dora. *Viés no aprendizado de máquina em sistemas de inteligência artificial: a diversidade de origens e os caminhos de mitigação*. Revista USP, São Paulo, Brazil, n. 135, p. 195–210, 2022. Available at: https://www.revistas.usp.br/revusp/article/view/206235. Accessed: 7 Jan. 2025.

in 2014[23] – is intrinsically linked to the perpetuation of discrimination.[24]

Thus, the origin of these biases is not restricted to data alone, as the authors explain. Although biases are frequently attributed to skewed databases, they can arise even before data collection, in the decisions of developers. If, for example, developers – even unconsciously – prioritize characteristics that favor a given group, the model will already be biased from its very conception.

From the perspective of the problem related to biases, it is extremely important to bring into the discussion the study of the heuristics of thought, which can generate errors in people's conclusions, carrying pre-defined patterns of thinking that do not translate into the best option for judgments, since in the great majority of cases they may be tied to premature judgments with deeply rooted prejudicial biases – a situation that is heightened when it comes to judicial decisions, which can significantly affect the lives of those involved.

---

[23] As reported by the British news agency Reuters, between 2014 and 2015 Amazon discovered that its new AI-assisted recruitment engine did not operate in a gender-neutral manner, such that the algorithm discriminated against women. The search for the best talent assigned ratings ranging from one to five stars; however, for software developer positions and other technical roles, the system considered female candidates inferior to male ones. The explanation was that the AI model had been trained on the database of résumés submitted to the company over a ten-year period, the vast majority of which came from men. Thus, Amazon's system taught itself that male candidates were preferable, discriminating against those that included "feminine" words. After the discovery, the tool was no longer used, and Amazon began to use a leaner version limited to helping with some basic tasks, such as removing duplicate applications from its database.

[24] According to Fabio Gagliardi Cozman and Dora Kaufman: "In general, biases are attributed entirely to biased databases. However, biases can emerge before data collection as a function of the decisions made by developers (the attributes and variables included in the model, indeed, determine the selection of the data). In the case of data-associated bias, there are two main origins: the collected data do not represent the proportional composition of the universe in question, or the data reflect the prejudices existing in society. The first case may occur, for example, if a training database contains more observations of a category that is in fact a minority; the second is illustrated by Amazon's automated recruitment-screening system deployed in 2014 (discontinued in early 2017). In 2015, the company identified that its system was not being gender-neutral, privileging male candidates. The system's AI algorithms were trained – that is, they 'learned' to identify patterns – on the database of résumés submitted to Amazon over a ten-year period, reflecting male dominance in the technology industry, that is, a biased reality (historical bias)."

Simplifications of thought can be divided into two systems of thinking (cognitive biases): the first is faster and more intuitive (cognitive shortcuts), while the second is characterized as more careful, which consequently affects its speed of conclusion, making it more costly. The first mode of reasoning, owing to its ease and lower effort requirement, may cause decisions that require deeper thought to be resolved more superficially, generating errors in their conclusions.[25]

Regarding dual-process theories, one cannot fail to mention the study by Daniel Kahneman,[26] *Thinking, Fast and Slow*, winner of the Nobel Prize in Economics, which describes how our thinking operates in two distinct yet interconnected systems. For the author, System 1, described above as intuitive and fast, is always active and can generate quick and efficient responses, but it can also lead to errors and biases. System 2, deliberative and slow, is activated when necessary and can correct the errors of System 1, but it demands more time and energy.

Daniel Boeing and Alexandre Rosa[27] explain that System 1 is fast, intuitive, and automatic, ideal for decisions requiring little effort and information. Essential for survival, it works through mental shortcuts (heuristics), which use past experiences to provide quick responses to familiar situations. However, its weakness is the pursuit of coherence at any cost, often confusing familiarity with truth and ignoring the quality of information, which can lead to systematic errors. In contrast, System 2 is slow, analytical, and deliberate. It requires focus and energy for complex tasks, such as abstract reasoning and the learning of new skills.[28]

---

[25] NUNES, Dierle; SANTOS, Natanael Lud; PEDRON, Flávio Quinaud. *Desconfiando da (Im)Parcialidade dos Sujeitos Processuais: Um estudo sobre os vieses cognitivos, o ruído, a mitigação de seus efeitos e o debiasing*. 3rd ed. São Paulo: Editora JusPodivm, 2022, p. 59.

[26] KAHNEMAN, Daniel. *Rápido e devagar: duas formas de pensar*. Translated by Cássio de Arantes Leite. Rio de Janeiro: Objetiva, 2012.

[27] BOEING, Daniel Henrique Arruda; ROSA, Alexandre Morais da. *Ensinando um robô a julgar: pragmática, discricionariedade, heurísticas e vieses no uso de aprendizado de máquina no judiciário*. 1st ed. Florianópolis: Academia, 2020, p. 81–82.

[28] Daniel Boeing and Alexandre Rosa: "System 1 serves to make quick decisions, with little cognitive effort and when little information is available, being commonly associated with intuitive or 'innate' processes. Especially in primitive societies, in which a quick response to an environmental stimulus could mean the difference between life and death, it was (and continues to be) quite useful. However, failing to take into account the influence of System 1 in decision-making pro-

The two systems cooperate: tasks that begin in System 2, such as driving, become automatic and migrate to System 1 with practice. The danger arises when, to avoid the effort of System 2, we simply accept the quick conclusions of System 1 on complex matters. We believe we are using reason (System 2), but in reality we are on autopilot, which results in flawed decisions based on biases and simplifications.

In the judicial sphere, this dynamic is critical. Case overload encourages judges and lawyers to use these mental shortcuts. Heuristics serve as "ready-made decisions" to reduce the workload and create a sense of coherence, but they risk oversim-

cesses can lead to systematic errors of thought. It operates through cognitive shortcuts, also called heuristics, which are nothing more than information-recognition mechanisms that help find simple – albeit more imperfect – answers to questions that would demand greater reflection and cognitive effort, characteristics of System 2. Thus, starting from an accumulation of past information, System 1 responds automatically when a similar situation occurs. In fact, this system is in search of coherence, so that it not infrequently ends up confusing 'familiarity and cognitive comfort with illusions of truth,' becoming insensitive 'to the quantity and quality of the information at the genesis of the impressions and intuitions.' In this way, the search for consistency in the stories narrated by System 1 is often generated without taking into account the deficiency of the available information. System 2, on the other hand, requires effort and attention and is related to logical, sequential, and conscious processes, but also slow and inefficient ones. Moreover, it is the only one that allows abstract thinking and 'programming the memory to obey instructions that embody non-habitual tasks,' such as finding objects with deliberately chosen characteristics among many others. In addition, the repetition of tasks that initially demanded System 2, such as driving or practicing a sport, ends up transferring them to System 1, allowing their automatic execution. The two systems, therefore, do not compete but cooperate, since there are mental processes that require System 1 and others that require System 2. However, owing to the effort demanded by System 2 and the speed and effectiveness of System 1, it is possible that, even in the face of complex questions, the former merely endorses the generalizations of the latter. When such an inversion occurs, the results can be disastrous, since one starts from exaggerated simplifications that lead to suboptimal decisions. Heuristics and biases therefore occur when System 1 is used while one believes System 2 is being used. This makes it evident how such cognitive failures can interfere in the judicial sphere. Especially in a situation of excessive demand, judges and other procedural subjects have neither the conditions nor the incentives to analyze in detail all the peculiarities of the cases that reach them, so that they create 'mental shortcuts' to decide. Such shortcuts or heuristics constitute 'ready-made decision mechanisms' and serve not only to reduce their mental workload but also to generate sensations of coherence and cognitive comfort" (BOEING, Daniel Henrique Arruda; ROSA, Alexandre Morais da. *Ensinando um robô a julgar…* 1st ed. Florianópolis: Academia, 2020, p. 81–82).

plifying complex cases, leading to suboptimal judgments.

Heuristics, as a mechanism intrinsic to human thought, permeate not only everyday decisions but also the very research in question and its interpretation. The influence of heuristics on legal research manifests itself in the selection of cases, the formulation of hypotheses, and, above all, the interpretation of results – the focal point of the present work.

In this respect, confirmation bias seeks to find information that corroborates or confirms a pre-established understanding, being the most tendentious bias to be adopted by the adjudicator. In this type of scenario, as a rule, the adjudicator selects the source from which the information is drawn in such a way that it may serve as an endorsement of their thinking, or selects ambiguous information that can be interpreted in whatever manner is most convenient to them.[29]

3.3 Automation Bias and the Anchoring Effect

The interaction between the reviewing judge and the draft decision generated by artificial intelligence does not occur in a rational vacuum. It is mediated by cognitive biases inherent in the human decision-making process. The two most prominent in this scenario are automation bias and the anchoring effect, which operate in synergy to potentially compromise the quality and independence of the supervision.

Automation bias describes the human tendency to overvalue and place excessive trust in the information and solutions proposed by automated systems, even when there is contradictory evidence. It manifests itself in two main forms: (i) when the human actively accepts an incorrect suggestion from the system – in our case, the judge accepts a flawed conclusion or reasoning proposed by the artificial intelligence simply because "the system generated it," presuming its correctness; and/or (ii) when the human fails to act or to perceive a problem because the system did not flag it. If the artificial intelligence, for example, fails to identify crucial evidence or a relevant line of argument, the human reviewer may also fail to notice it, trusting that the

[29] NUNES, Dierle; SANTOS, Natanael Lud; PEDRON, Flávio Quinaud. *Desconfiando da (Im)Parcialidade dos Sujeitos Processuais…* 3rd ed. São Paulo: Editora JusPodivm, 2022, p. 81.

algorithm's "fine-tooth comb" was complete.[30]

The logic behind this bias is a mixture of the perception of the machine's superiority (it is seen as faster, more objective, and with access to more data) and a strategy of conserving mental effort. It is cognitively less costly to trust than to verify.

The anchoring effect, in turn, can be defined as an individual's tendency to rely heavily on the first piece of information received (the "anchor") when making decisions.[31] Subsequent estimates and judgments are adjusted from that anchor, and such adjustments are frequently insufficient.[32] The following scenario may illustrate this: the first proposal, drafted exclusively by an artificial intelligence system, establishes a starting point on the adjudicator's mental map. Even if the judge disagrees and tries to move away from the anchor, the final decision tends to remain closer to it than it would have been had the process begun "from scratch."

Applied to the scenario of reviewing a judgment, these two biases not only coexist – they reinforce each other. First, because the draft judgment generated by the artificial intelligence is the first and most structured piece of information the judge analyzes. It functions as an extremely powerful anchor, defining not only whether or not to grant the claim, but also the narrative structure, the selection of facts, the choice of evidence to be valued, and the line of legal argument – that is, it becomes the "anchor" for subsequent cases.

From this prior scenario, automation bias emerges as a means of reinforcing the aforementioned "anchor" created by the judge. Imagine the hypothetical scenario of use in which the adjudicator might think: "This decision was drafted by an artificial intelligence system that analyzed a far greater quantity of decisions than a human could, so the probability that it is correct is much higher than my own understand-

---

30 TAVARES, João Paulo Lordelo Guimarães. *Vieses implícitos e técnicas de automação decisória: riscos e benefícios*. Civil Procedure Review, v. 12, n. 1, p. 105–132, 2021. Available at: https://www.civilprocedurereview.com/revista/article/view/225. Accessed: 2 July 2025.

31 TVERSKY, Amos; KAHNEMAN, Daniel. *Judgment under uncertainty: Heuristics and biases*. Science, v. 185, n. 4157, 1974.

32 LIMA, George. *A âncora da justiça: o efeito de ancoragem nas decisões judiciais*. Civilistica.com, Rio de Janeiro, v. 10, n. 3, p. 1–23, 2021. Available at: https://civilistica.emnuvens.com.br/redc/article/view/569. Accessed: 2 July 2025.

ing." This (mis)placed confidence reduces the motivation to depart significantly from the anchor.

The practical result of this entire scenario is that the "review" becomes a mere editing exercise – changing words and validating – rather than an act of genuine review and, consequently, an independent judgment. One can imagine practical situations such as, in the analysis of evidence, where the judge may be led to overvalue the evidence that the artificial intelligence highlighted in its draft and to underestimate or ignore evidence that the algorithm disregarded; or, in the legal reasoning, where the review may be limited to checking whether the case law cited by the artificial intelligence is applicable, rather than questioning whether it was the best or the only viable legal thesis. Creativity and the search for alternative legal solutions are inhibited by the anchor. Finally, in this scenario a false sense of authorship is created. By making small corrections to the text – adjusting a comma, replacing a synonym, adding a sentence – the adjudicator understands that they have taken ownership of the piece and are its true author. However, the logical structure and the final result, defined by the anchor, remain intact. The review becomes a "legitimation" of the machine's proposal, not an autonomous judgment.

In this way, any regulation that genuinely seeks to mitigate the risks mentioned in the present study must go beyond the simple requirement of "human supervision." It must provide mechanisms to actively mitigate these biases – such as training programs on cognitive psychology for judges, or the development of artificial intelligence systems that, instead of presenting a single draft, offer alternative scenarios, highlighting points of uncertainty and forcing the adjudicator to make critical choices rather than merely saying "yes" to a ready-made solution.

4. CONCLUSION

Based on the research carried out, it is concluded that the requirement of human supervision in the artificial intelligence systems used in the Judiciary does not hold up as a technically and legally effective safeguard. Although provided for as a central element of algorithmic governance, the presence of a "human reviewer" is not, by itself, capable of ensuring effective control over automated decisions, nor of guaranteeing full observance of due process of law.

This conclusion stems from the identification of three central structural weaknesses. First, it was observed that the foundation models adopted at scale – such as large language models – operate opaquely, with non-interpretable statistical dynamics that escape causal explanation. Such technical opacity renders substantive human review unfeasible, even where it is formally provided for. Second, it was found that the human adjudicator, when interacting with the machine, is exposed to relevant cognitive biases, such as automation bias and the anchoring effect, which reduce their capacity to exercise independent and critical control over the results suggested by the system. Finally, the Resolution suffers from an absence of normative differentiation between low- and high-risk AI applications, applying generic requirements to diverse contexts, which compromises both regulatory efficiency and the protection of fundamental rights.

On the basis of these elements, it is found that the figure of the human reviewer, in the form in which it is currently regulated, acts more as a symbolic instrument of legitimacy than as a functional mechanism of accountability and control. By reinforcing the appearance of supervision without providing the effective means for it to occur in a critical and informed manner, the current model incurs a regulatory fallacy that masks the technical limitations of the systems and transfers, in a scarcely transparent manner, the responsibility for automated decisions to the human agent.

REFERENCES

BALDWIN, Robert; CAVE, Martin; LODGE, Martin. *Understanding Regulation: Theory, Strategy, and Practice*. Oxford: Oxford University Press, 2012.

BOEING, Daniel Henrique Arruda; ROSA, Alexandre Morais da. *Ensinando um robô a julgar: pragmática, discricionariedade, heurísticas e vieses no uso de aprendizado de máquina no judiciário*. 1st ed. Florianópolis: Academia, 2020.

BRASIL. Conselho Nacional de Justiça. *Resolução nº 615, de 18 de março de 2025*. Provides for the governance, development, and use of artificial intelligence in the Judiciary. Diário da Justiça Eletrônico, Brasília, DF, n. 92, 20 Mar. 2025.

CASPER, Stephen et al. *Black-box access is insufficient for rigorous AI audits*. In: THE 2024 ACM CONFERENCE ON FAIRNESS, ACCOUNTABILITY, AND TRANSPARENCY, 2024, Rio de Janeiro. New York: ACM, 2024. Available at: http://dx.doi.org/10.1145/3630106.3659037. Accessed: 19 June 2025.

CONSELHO NACIONAL DE JUSTIÇA (CNJ). *Justiça em Números 2024*. Portal CNJ. Brasília, 2024, p. 133. Available at: https://www.cnj.jus.br/pesquisas-judiciarias/justica-em-numeros/. Accessed: 3 June 2025.

COZMAN, Fabio Gagliardi; KAUFMAN, Dora. *Viés no aprendizado de máquina em sistemas de inteligência artificial: a diversidade de origens e os caminhos de mitigação*. Revista USP, São Paulo, Brazil, n. 135, p. 195–210, 2022. Available at: https://www.revistas.usp.br/revusp/article/view/206235. Accessed: 7 Jan. 2025.

DISTRITO FEDERAL. Tribunal de Justiça do Distrito Federal e dos Territórios. 2nd Special Civil Court of Brasília. *Case No. 0700031-85.2025.8.07.0016*. Judgment. Claimant: Regis Stanley Mattioli. Defendant: 99 Tecnologia Ltda. Judge: Flávio Augusto Martins Leite. Brasília, 25 Mar. 2025.

G'SELL, Florence. *Regulating under Uncertainty: Governance Options for Generative AI*. SSRN Electronic Journal, 6 Aug. 2024. Available at: http://dx.doi.org/10.2139/ssrn.4918704. Accessed: 19 June 2025.

INTERNATIONAL BUSINESS MACHINES CORPORATION – IBM. *What is algorithmic bias?* Available at: https://www.ibm.com/br-pt/think/topics/algorithmic-bias. Accessed: 3 June 2025.

JUDGE, Brian; NITZBERG, Mark; RUSSELL, Stuart. *When code isn't law: rethinking regulation for artificial intelligence*. Policy and Society. Oxford: Oxford University Press, p. 1–13, 2024.

KAHNEMAN, Daniel. *Rápido e devagar: duas formas de pensar*. Translated by Cássio de Arantes Leite. Rio de Janeiro: Objetiva, 2012.

LESSIG, Lawrence. *Code is law: on liberty in cyberspace*. Harvard Magazine, Cambridge, Jan./Feb. 2000, p. 1.

LIMA, George. *A âncora da justiça: o efeito de ancoragem nas decisões judiciais*. Civilistica.com, Rio de Janeiro, v. 10, n. 3, p. 1–23, 2021. Available at: https://civilistica.emnuvens.com.br/redc/article/view/569. Accessed: 2 July 2025.

LOEVINGER, Lee. *Jurimetrics: the methodology of legal inquiry*. New York: Basic Books, 1963, p. 6.

MENG, Xiao-Li. *Data science and engineering with human in the loop, behind the loop, and above the loop*. Harvard Data Science Review, v. 5, n. 2, Spring 2023. Available at: https://hdsr.mitpress.mit.edu/pub/812vijgg/release/3. Accessed: 19 June 2025.

NARAYANAN, Arvind; KAPOOR, Sayash. *AI Snake Oil: What Artificial Intelligence Can Do, What It Can't, and How to Tell the Difference*. Princeton: Princeton University Press, 2024.

NUNES, Dierle; SANTOS, Natanael Lud; PEDRON, Flávio Quinaud. *Desconfiando da (Im)Parcialidade dos Sujeitos Processuais: Um estudo sobre os vieses cognitivos, o ruído, a mitigação de seus efeitos e o debiasing*. 3rd ed., rev., updated, and expanded. São Paulo: Editora JusPodivm, 2022.

RUDIN, Cynthia. *Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead*. Nature Machine Intelligence, v. 1, n. 5, p. 206, 2019. Available at: https://doi.org/10.1038/s42256-019-0048-x. Accessed: 19 June 2025.

SCHWEDE, Matheus Antes. *Inteligência Artificial Tendenciosa: Discriminação racial pelo viés algorítmico*. Londrina, PR: Thoth, 2023.

SUSSKIND, Richard. *Tomorrow's Lawyers: An Introduction to Your Future*. 2nd ed. Oxford: Oxford University Press, 2017. 240 p.

TAVARES, João Paulo Lordelo Guimarães. *Vieses implícitos e técnicas de automação decisória: riscos e benefícios*. Civil Procedure Review, v. 12, n. 1, p. 105–132, 2021. Available at: https://www.civilprocedurereview.com/revista/article/view/225. Accessed: 2 July 2025.

TVERSKY, Amos; KAHNEMAN, Daniel. *Judgment under uncertainty: Heuristics and biases*. Science, v. 185, n. 4157, 1974.

SECTORIAL REGULATION OF ARTIFICIAL INTELLIGENCE IN ELECTIONS: THE CASE OF BRAZIL

Ana Luiza Marques

Ph.D. Candidate and Master of Laws, Federal University of Minas Gerais (UFMG), Brazil. Visiting Researcher, Dublin City University (DCU), Ireland. Lawyer.

DOI: https://doi.org/10.59224/dti8.ch24

Abstract: This article analyzes Brazilian Superior Electoral Court (TSE) Resolution nº 23.732/2024 as an example of sectorial AI regulation. It argues that electoral processes are high-risk contexts and that sectorial approaches offer a more flexible, context-sensitive regulatory strategy. The first section contrasts general and sectorial models of AI regulation. The second examines transnational challenges of AI use in elections and the need for specific legal safeguards. The third explores Brazil's regulatory response through Resolution nº 23.732/2024, in comparison with other jurisdictions. The final section offers a critical assessment of the Resolution's strengths and shortcomings. While it marks a significant step toward AI governance in electoral contexts, broader frameworks and stronger institutional mechanisms are required to ensure effective regulation.

Keywords: Artificial Intelligence Regulation – Electoral Integrity – Sectorial Governance

Table of contents: Introduction – 1. Artificial Intelligence Regulation: General Approaches and Sectorial Strategies – 2. 2. Social media platforms and artificial intelligence: transnational challenges to democracy and elections – 3. The Brazilian Response: Resolution nº 23.732/2024 from the Superior Electoral Court – 3.1. Contextual Background: Electoral Institutions and Disinformation Challenges – 3.2.

Resolution nº 23.732/2024: Regulatory Innovations and Enforcement Mechanisms – 3.3. Comparative Perspectives: International Approaches to AI and Electoral Integrity – Conclusion

INTRODUCTION

Artificial intelligence (AI) has introduced complex challenges to the integrity of electoral processes worldwide. While it offers potential benefits in automating tasks and enhancing communication,[1] its use in political campaigns – especially through generative models – raises serious concerns about manipulation, disinformation, and the erosion of public trust in democratic institutions.[2] In this context, elections emerge as a particularly high-risk domain that demands targeted regulatory strategies.

This article examines, through a comparative lens, how Brazil has responded to these challenges through a sector-specific initiative: Resolution nº 23.732/2024, enacted by the Brazilian Superior Electoral Court (TSE). The Resolution addresses the use of AI in electoral campaigns, establishing legal obligations for candidates and application providers. The paper addresses the following questions: (i) what legal and democratic challenges arise from the use of AI in elections? (ii) how does Resolution nº 23.732/2024 address these challenges? (iii) to what extent can this sectorial approach serve as a model for other countries and to the broader debate on AI governance?

To answer these questions, the article adopts legal and comparative analysis. It

---

1 For example, AI has the latent ability to enhance election administration, aid in detecting and preventing fraud, improve data management, and strengthen electoral security and cyber defense strategies. See: International Institute for Democracy and Electoral Assistance. *Artificial Intelligence for Electoral Management: A Guide for Electoral Management Bodies.* International IDEA, 2024. https://www.idea.int/sites/default/files/2024-04/artificial-intelligence-for-electoral-management.pdf UNESCO. *Elections in Digital Times: A Guide for Electoral Practitioners.* Paris: UNESCO, 2023. https://unesdoc.unesco.org/ark:/48223/pf0000382102

2 In a message for Sunday's International Day of Democracy, the United Nations (UN) Secretary-General, António Guterres, warned about the risks of unchecked AI in electoral processes. United Nations. UN Warns of Disinformation Threat to 2024 Electoral Cycle, Global Stability. *UN News*, September 19, 2024. https://news.un.org/en/story/2024/09/1154316.

draws on a literature review, a close examination of Resolution nº 23.732/2024, and selected international regulatory experiences concerning the use of AI in elections. The comparative dimension highlights divergences in normative scope, institutional design, and enforcement capacity. Throughout the drafting process of this article, AI tools – particularly ChatGPT – were employed to support the revision and refinement of the text, under the author's supervision and critical judgment, in line with academic integrity standards.[3]

To structure this discussion, the article initially presents an overview of current debates on AI regulation, focusing on the distinction between general regulatory frameworks and sectorial strategies. It explores the concept of "sectorial AI regulation" as a more effective strategy for addressing risks in specific applications such as healthcare, public safety, and elections, as it offers greater flexibility and alignment with concrete risks, in contrast to overly generic regulatory models. Nevertheless, a general regulatory framework that establishes overarching principles and guidelines for the broad use of AI remains important, particularly to guide and harmonize sectorial norms.

The article then turns to the role of digital platforms – as central spaces for the dissemination of information and political debate – and AI in electoral campaigns. It examines how these technologies, particularly generative AI systems, amplify long-standing challenges related to disinformation, manipulation, polarization and the distortion of public debate. Given the speed, scale, and opacity with which AI-generated content can spread, elections must be recognized as a high-risk context that demands specific legal and institutional safeguards capable of preserving electoral integrity and democratic resilience.

The third section of the paper examines the TSE's Resolution nº 23.732/2024, situating it within Brazil's legal landscape. It begins by contextualizing the institutional role of the Electoral Justice and the normative competence of the TSE, highlighting how this has enabled the Court to adopt a sectorial approach to regulating

[3] This approach follows current international academic integrity guidelines, such as those issued by COPE (2023) and major scientific journals, which recommend transparency and human accountability when using AI tools for academic purposes. See: COPE Council. COPE position - Authorship and AI - English. https://doi.org/10.24318/cCVRZBms

the use of AI in elections. It then analyzes the content of the Resolution, which introduces both prohibitive and preventive measures: it bans deepfakes in political advertising, requires the labeling of synthetic media, restricts the use of chatbots and avatars, and imposes obligations on digital platforms to detect and remove disinformation. Next, it presents a comparative analysis with regulatory efforts in other jurisdictions – such as South Korea, Singapore, the United States, China, and the European Union – and evaluates how these approaches differ in scope, enforceability, and focus. Finally, it explores the challenges faced by the Brazilian model, including the absence of a general AI law, the dependence on judicial enforcement, and emerging tensions between institutional ambition and implementation capacity.

Taken together, the analysis concludes that Resolution nº 23.732/2024 represents an innovative and context-sensitive approach to AI governance in elections. By articulating obligations for both political actors and platform providers, the Brazilian model offers important lessons for other jurisdictions navigating the intersection of technology, democracy, and electoral integrity. However, its effectiveness will ultimately depend on the development of a broader regulatory ecosystem – including a general AI law – and on the construction of institutional capacity, as well as a consistent jurisprudence capable of ensuring legal certainty, coherence, and accountability in the face of evolving technological risks.

1. ARTIFICIAL INTELLIGENCE REGULATION: GENERAL APPROACHES AND SECTORIAL STRATEGIES

The rapid development of AI capabilities – especially with generative AI and large language models (LLMs) – and the increased use of AI-based systems will likely play a transformative role in society, with both positive and negative effects. Despite its enormous potential benefits,[4] the negative side effects of highly capable AI systems represent a significant risk to society and, more deeply, to vulnerable groups. Ac-

[4] AI has, for example, the potential to transform labor markets, improve access to justice and contribute to economic growth. ARBEL, Yonathan A.; TOKSON, Matthew J.; LIN, Albert. Systemic Regulation of Artificial Intelligence (December 16, 2023). *AI Safety Legal Paper Series* 12-24, 56 Ariz. St. L. J. 545, University of Utah College of Law Research Paper No. 582, Available at SSRN: https://ssrn.com/abstract=4666854

cording to the literature, those risks are mostly related to (i) the scale of AI systems; (ii) the unpredictable nature of these systems; and (iii) the type and extent of the harms caused by AI.[5]

Regarding the first point, the large-scale nature of many AI systems – notably foundation models –[6] amplifies their potential impact, allowing both benefits and harms to spread rapidly and widely across different sectors of society. As for the second point, the unpredictability is usually related to the lack of interpretability and explainability,[7] as AI systems are "*built on black-box data driven systems*",[8] meaning that the "*systems' behavior is an emergent property. (…) its behavior is neither expressed intentionally by designers in software program code nor legible (yet) by examining the program code*".[9] Finally, the third point concerns the nature and extent of

---

5 Regarding the present harms of AI, authors usually highlight harms related to bias and discrimination; fraud; privacy; social manipulation and surveillance. It is also possible to mention harms that concern the use of AI in wars (autonomous weapons, for example), job displacement and negative impacts on the environment and also the increased dependence on AI. See: ARBEL, Yonathan A.; TOKSON, Matthew J.; LIN, Albert. Systemic Regulation of Artificial Intelligence (December 16, 2023). *AI Safety Legal Paper Series* 12-24, 56 Ariz. St. L. J. 545, University of Utah College of Law Research Paper No. 582, Available at SSRN: https://ssrn.com/abstract=4666854. CAJUEIRO, Daniel O.; CELESTINO, Victor R. R. A *Comprehensive Review of Artificial Intelligence Regulation:* Weighing Ethical Principles and Innovation. 09.09.2024. Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4950727.

6 Foundation model is a "model that is trained on broad data (generally using self-supervision at scale) that can be adapted (e.g., fine-tuned) to a wide range of downstream tasks (…) From a technological point of view, foundation models are not new — they are based on deep neural networks and self-supervised learning (…) However, the sheer scale and scope of foundation models from the last few years have stretched our imagination of what is possible; for example, GPT-3 has 175 billion parameters and can be adapted via natural language prompts to do a passable job on a wide range of tasks despite not being trained explicitly to do many of those tasks". BOMMASANI, R. et al. On the opportunities and risks of foundation models. 2021. ArXiv:2108.07258.

7 CELESTINO, Victor R. R. A *Comprehensive Review of Artificial Intelligence Regulation:* Weighing Ethical Principles and Innovation. 09.09.2024. Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4950727.

8 JUDGE, Brian; NITZBERG, Mark; RUSSELL, Stuart. When code isn't law: rethinking regulation for artificial intelligence, *Policy and Society*, Volume 44, Issue 1, January 2025, p. 2.

9 JUDGE, Brian; NITZBERG, Mark; RUSSELL, Stuart. When code isn't law: rethinking regulation

harms that may result from the use of AI systems. The literature often highlights harms related to bias and discrimination, fraud and misinformation, and violations of privacy.[10] It is also possible to identify broader and more systemic risks, including social manipulation and surveillance, the use of AI in warfare (e.g., autonomous weapons), job displacement, large-scale environmental impacts, and increased human dependence on AI technologies.[11]

The novel challenges of AI – notably generative AI – therefore call for the development of a regulatory approach capable of addressing the specific risks associated with its use, as the costs to society (externalities), information asymmetries, threats to human rights, and the erosion of social solidarity all serve as compelling justifications for regulatory intervention.[12] Even though technical solutions are indeed capable of addressing part of the risks associated with generative AI (e.g., technical ideas to reduce the fabrication of citations and information), law is still needed to mitigate most of the issues and enhance the positive outcomes,[13] since "*technical solutions are part of the puzzle but so are the underlying values and priorities of society and the*

for artificial intelligence, *Policy and Society*, Volume 44, Issue 1, January 2025, p. 2. See also: ARBEL, Yonathan A.; TOKSON, Matthew J.; LIN, Albert. Systemic Regulation of Artificial Intelligence (December 16, 2023). *AI Safety Legal Paper Series* 12-24, 56 Ariz. St. L. J. 545, University of Utah College of Law Research Paper No. 582, Available at SSRN: https://ssrn.com/abstract=4666854.

10 ARBEL, Yonathan A.; TOKSON, Matthew J.; LIN, Albert. Systemic Regulation of Artificial Intelligence (December 16, 2023). *AI Safety Legal Paper Series* 12-24, 56 Ariz. St. L. J. 545, University of Utah College of Law Research Paper No. 582, Available at SSRN: https://ssrn.com/abstract=4666854. CELESTINO, Victor R. R. A *Comprehensive Review of Artificial Intelligence Regulation:* Weighing Ethical Principles and Innovation. 09.09.2024. Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4950727.

11 JUDGE, Brian; NITZBERG, Mark; RUSSELL, Stuart. When code isn't law: rethinking regulation for artificial intelligence, *Policy and Society*, Volume 44, Issue 1, January 2025, p. 2-3.

12 BALDWIN, Robert; CAVE, Martin; LODGE, Martin. *Understanding Regulation*: Theory, Strategy, and Practice. 2nd ed. Oxford: Oxford University Press, 2012, chapter 2.

13 "In the long run, we think that a combination of technology and law can alleviate most of the problems we've described, as well as amplify the benefits". NARAYANAN, Arvind; KAPOOR, Sayash. *AI Snake Oil:* What Artificial Intelligence Can Do, What It Can't, and How to Tell the Difference. Princeton, NJ: Princeton University Press, 2024, p. 8.

capacity of nation states to implement real solutions".[14]

Even so, acknowledging the need for regulation is not enough. Risk mitigation in the field of AI will not occur "by default".[15] Nevertheless, there is still no consensus on how AI should be regulated. As a result, "*nearly everyone agrees government regulation of generative AI is necessary, but there is little consensus around the form it should take*".[16]

In this context, this paper focuses on two main regulatory approaches that have gained prominence: general regulatory frameworks and sectorial strategies.[17] General regulatory frameworks aim to establish overarching principles and safeguards applicable to all AI systems, regardless of their context of use. The EU AI Act is an example of this model, as it adopts a horizontal approach that provides "general provisions setting out an overall framework, a reference context within which artificial intelligence systems operate, both today and in the future".[18] To that end, the EU AI Act adopts a risk-based regulatory model, under which AI systems are classified into different categories according to the level of risk they pose. Systems deemed to pose unacceptable risks are prohibited, while high-risk systems are subject to strict ex ante conformity assessments and robust requirements regarding traceability, transparency, human oversight, accuracy, robustness, and cybersecurity. By contrast, low-risk and minimal-risk systems are subject to lighter obligations, relying

---

[14] JUDGE, Brian; NITZBERG, Mark; RUSSELL, Stuart. When code isn't law: rethinking regulation for artificial intelligence, *Policy and Society*, Volume 44, Issue 1, January 2025, p. 10.

[15] "AI holds the potential for both unprecedented benefits and unprecedented harms. Significant efforts are required to ensure that AI's increasing impact on society is beneficial Regulators cannot assume that this will happen by default." JUDGE, Brian; NITZBERG, Mark; RUSSELL, Stuart. When code isn't law: rethinking regulation for artificial intelligence, *Policy and Society*, Volume 44, Issue 1, January 2025, p. 9-10.

[16] JUDGE, Brian; NITZBERG, Mark; RUSSELL, Stuart. When code isn't law: rethinking regulation for artificial intelligence, *Policy and Society*, Volume 44, Issue 1, January 2025, p. 2.

[17] FINOCCHIARO, Giusella. The regulation of artificial intelligence. *AI & Soc 39*, 1961–1968 (2024). https://doi.org/10.1007/s00146-023-01650-z. ELLUL, Joshua. Should we regulate AI or some uses of software. *Discov Artif Intell* **2**, 5 (2022). https://doi.org/10.1007/s44163-022-00021-9

[18] FINOCCHIARO, Giusella. The regulation of artificial intelligence. *AI & Soc 39*, 1961–1968 (2024). https://doi.org/10.1007/s00146-023-01650-z

instead on voluntary codes of conduct or transparency measures.[19]

General frameworks can promote coherence, legal certainty, and international harmonization (as, for example, in the case of OECD AI Principles).[20] However, they often struggle to respond effectively to the technical and social specificities of certain high-risk applications in specific sectors and for specific uses, leading to concerns about abstraction and regulatory gaps, as well as the risk of regulatory overlap or unnecessary burdens.[21] A technology-wide risk-based model may impose obligations on systems that clearly do not require such regulation. For example, the level of transparency and human oversight required for LLMs used to summarize judicial rulings is not comparable to that required for LLMs used to assist or make judicial decisions, which directly affect fundamental rights. Applying identical regulatory burdens to both may result in inefficiencies and fail to reflect the different risk profiles involved in each specific application.

In contrast, sectorial regulation seeks to tailor legal obligations to the specific risks and characteristics of AI applications within particular domains, such as healthcare, education, or elections. By taking into account the context of use, the nature of the technology, and the potential for harm, this approach allows for more proportionate,

---

[19] See: GIKAY ,Asress Adimi. Risks, innovation, and adaptability in the UK's incrementalism versus the European Union's comprehensive artificial intelligence regulation, *International Journal of Law and Information Technology*, Volume 32, 2024, eaae013, https://doi.org/10.1093/ijlit/eaae013. FINOCCHIARO, Giusella. The regulation of artificial intelligence. *AI & Soc 39*, 1961–1968 (2024). https://doi.org/10.1007/s00146-023-01650-z EBERS, Martin. *Truly Risk-Based Regulation of Artificial Intelligence:* How to Implement the EU's AI Act. 19.06.2024. Available at: <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4870387>

[20] The OECD AI principles were the first intergovernmental standard on AI, adopted by the OECD Council at Ministerial level on 22 May 2019. OECD. *Recommendation of the Council on Artificial Intelligence*. OECD/LEGAL/0449, 2019. Available at: https://oecd.ai/en/assets/files/OECD-LEGAL-0449-en.pdf>.

[21] According to Joshua Ellul, "The trade-off for a technology-wide risk-based approach is that determining an application's risk level on all systems may impose overheads on systems that would clearly not require regulation, and that it may result in duplication of laws where such regulation is already imposed on technological solutions (whether directly or indirectly)". ELLUL, Joshua. Should we regulate AI or some uses of software. *Discov Artif Intell* **2**, 5 (2022). https://doi.org/10.1007/s44163-022-00021-9

effective, and context-sensitive rules that focus on the specific uses of the technology. For instance, the use of AI in criminal sentencing or medical diagnosis justifies stricter safeguards than its use in administrative tasks or informational services within the same sectors. It also enables regulators to define targeted requirements for safety, transparency, and accountability, aligned with the societal and ethical concerns specific to each field. That is the prevalent model in the United States, as regulatory authority is distributed among various federal agencies, each responsible for overseeing AI applications within its respective domain.[22]

However, sectorial regulation may also lead to fragmentation, overlapping mandates, and inconsistencies across different regulatory regimes, especially in the absence of guiding general principles or coordination mechanisms. Thus, a third approach, that unites elements of both general (horizontal) and sectorial (vertical) regulation seems to be the best solution to assure (i) the flexibility and alignment with concrete risks offered by sectorial regulation, as opposed to overly generic models; (ii) while also establishing overarching principles and guidelines for the broad use of AI, to guide and harmonize sectorial norms.

These regulatory debates are not merely theoretical. They have practical consequences in sectors where AI systems are already being broadly used. One of the most complex and pressing domains is the governance of AI in the context of elections. The next section explores how the use of AI raises transnational challenges to democratic integrity, and how it is important to develop a consistent regulatory approach to address these new challenges.

2. SOCIAL MEDIA PLATFORMS AND ARTIFICIAL INTELLIGENCE: TRANSNATIONAL CHALLENGES TO DEMOCRACY AND ELECTIONS

In the digital era, algorithmically mediated platforms became the locus for economic, social, and even public activities, reshaping the way they are performed.[23] As

---

[22] NARAYANAN, Arvind; KAPOOR, Sayash. *AI Snake Oil:* What Artificial Intelligence Can Do, What It Can't, and How to Tell the Difference. Princeton, NJ: Princeton University Press, 2024, p. 270.

[23] According to Julie Cohen: "The emergence of platform-based business models has reshaped

our daily interactions are increasingly intermediated by these platforms,[24] they also detain the power to unilaterally settle the rules to regulate the activities developed in these environments. As a result, these companies now have a leading role "in underpinning and regulating our economic and social activities".[25]

This power is formally exercised through private agreements, such as terms of service and community guidelines. Nevertheless, as these companies perform public interest functions, they are substantially dealing with the scope of fundamental rights and freedoms (for example, freedom of expression, freedom to conduct a business, personality and privacy rights, due process, among others), and consequently with democratic (and constitutional) values.[26] This means that a small number of companies, usually referred as big techs or tech giants,[27] detain not only private economic powers, but also social and political ones, though they are not the result of a democratic deliberation process and not subject to the same accountabilities standards and constraints applied to state power.[28]

---

patterns of financial activity, workforce management, and land ownership. Platforms have also reshaped information transmission, entertainment, social interaction, and consumption of goods and services, and have destabilized the locally embedded systems that previously mediated those activities in many different types of communities". COHEN, Julie E. *Between truth and power*: The legal constructions of information capitalism. New York: Oxford University Press, 2019, p. 37.

[24] "According to one study, 70 percent of industries in the United States are now influenced by digital platforms (…). An estimated $490 billion's worth of goods passed through Amazon's marketplace in 2011— more than most countries' entire gross domestic product." LEHDONVIRTA, Vili. *Cloud empires: how digital platforms are overtaking the state and how we can regain control.* Cambridge, Massachusetts; London, England: The MIT Press, 2022, p. 3

[25] LEHDONVIRTA, Vili. *Cloud empires: how digital platforms are overtaking the state and how we can regain control.* Cambridge, Massachusetts; London, England: The MIT Press, 2022, p. 3

[26] GREGORIO, Giovanni de. *Digital Constitutionalism in Europe*: reframing rights and powers in the algorithmic society. Cambridge: Cambridge University Press, 2022, p. 17.

[27] Companies that usually fit into this label are Apple, Amazon, Microsoft, Google/Alphabet, and Facebook/Meta, but one can also mention the Chinese ones, such as Tencent, Alibaba, Baidu, ByteDance, Meituan e Didi Xuxing.

[28] GREGORIO, Giovanni de. *Digital Constitutionalism in Europe*: reframing rights and powers in the algorithmic society. Cambridge: Cambridge University Press, 2022, p. 18. See also: CELESTE, Edoardo, *Digital constitutionalism: the role of internet bills of rights*, Abingdon, Oxon; New York, NY: Routledge, 2021.

After years of operating freely and apart from the state, mostly in a regime of private self-regulation (at least in liberal States),[29] the risks involved in allowing such a huge concentration of power in the hands of a few companies have become increasingly evident.[30] Hence, discussions on internet governance gained momentum in global and regional levels, considering that the dangers arising from the consolidation of these transnational private powers bring new challenges to constitutional democracies, but also the risks that are not equally divided among developed and developing countries.[31] This context highlights the need not only for global govern-

[29] Illiberal countries, such as China, Saudi Arabia and Russia aim centralization, surveillance and a strong state control over the digital sphere. China long-term govern-driven national strategy, for its relevance in the global level, is an example of the distinction between the problems faced by liberal and illiberal countries in the digital realm. Internet shutdowns are one of the clearest examples of the harms presented by a strong state control over the digital sphere. See: HILLMAN, Jonathan E., *The digital Silk Road*: China's quest to wire the world and win the future, First Edition. New York: Harper Business, 2021. GREGORIO, Giovanni De; STREMLAU, Nicole, Internet Shutdowns in Africa | Internet Shutdowns and the Limits of Law, *International Journal of Communication,* v. 14, n. 0, p. 20, 2020.

[30] Some examples are waves of disinformation, hate speech, and attacks on democracy, abuses of market power, market segmentation, harms to users' mental health, social polarization, and the instrumentalization of this powerful informational infrastructure to achieve political aims. See: VAIDHYANATHAN, Siva. *Antisocial Media:* How Facebook Disconnects Us and Undermines Democracy. 1ª ed. Oxford University Press, 2018; MOROZOV, Evgeny. *Big Tech:* a ascensão dos dados e a morte da política. São Paulo: Ubu Editora, 2018; TUFEKCI, Zeynep. The road from Tahrir to Trump. *MIT Technology review*, vol. 121. n. 05, set/out 2018. KHAN, Lina. Amazon's Antitrust Paradox. *Yale Law Journal 126*, n. 3, jan. 2017, 710-805. COHEN, Julie E. *Between truth and power*: The legal constructions of information capitalism. New York: Oxford University Press, 2019, p. 70. HAO, Karen. How Facebook got addicted to spreading misinformation. *MIT Technology Review*, 11.03.2021. Available at: <https://www.technologyreview.com/2021/03/11/1020600/facebook-responsible-ai-misinformation/>. Access: 28.07.2023.

[31] Among the harmful consequences of this new form of production is the increase of informational disparities between companies and users, companies, and the states, but also between developed and developing countries, as the largest tech companies are located in few jurisdictions (mainly, the U.S. and China). Snowden's revelations regarding existing surveillance practices to track online communication by the US National Security Agency (NSA) are one example of the harms posed by these power and information asymmetries. Another example of the damaging effects of these disparities has surfaced with the Facebook Papers. They show that there are big asymmetries between efforts adopted by the platform in the Global North and the Global South, as well as a

ance strategies, but also for context-sensitive, regional responses.

The rise and popularization of generative AI – which has become widely accessible – adds a new layer of complexity to this regulatory landscape. As AI-generated content becomes increasingly sophisticated and emotionally persuasive, there is a growing global concern about the misuse of this technology in elections, especially considering the transnational character of the digital revolution. The main risks associated with the use of AI in elections include the intensification of echo chambers, political polarization and social manipulation,[32] all of which threaten democratic processes and undermine trust in electoral institutions. As Yonathan Arbel, Matthew Tokson and Albert Lin point out:

> *Democracies are built around systems of shared trust and governance. Voting requires individuals to believe that their votes matter, that the information people receive is-at least generally-accurate, and that the elections are legitimately run. Absent those, the very democratic compromise is jeopardized.*
>
> *Future Al systems may strain assumptions of trust. Deepfakes and voice cloning are becoming increasingly persuasive, making it difficult to verify whether a statement is given by a politician or a fraudster. AI-generated misinformation is currently as effective, or even more so, than the human-generated kind-and it is much easier to produce in massive quantities. (…)*
>
> *Election interference, in the form of astroturfing, misinformation pollution, or other social engagement, will likely also rise in effectiveness. Using an LLM*

lack of interest in adopting measures aimed at combating disinformation in countries where there is less risk of regulation. See: GREENWALD, Glenn. *No place to hide*: Edward Snowden, the NSA, and the Surveillance State. New York, NY: Metropolitan Books, 2014, p. 77-146. LIMA, Cristiano. A whistleblower's power: Key takeaways from the Facebook Papers. *The Washington Post,* 26.10.2021. https://www.washingtonpost.com/technology/2021/10/25/what-are-the-facebook-papers/. MELLO, Patrícia Campos. Facebook não prioriza Brasil contra fake news, diz Frances Haugen, do Facebook Papers. *Folha de São Paulo*, 03.07.2022. https://www1.folha.uol.com.br/poder/2022/07/facebook-nao-prioriza-brasil-contra-fake-news-diz-frances-haugen-do-facebook-papers.shtml

[32] CAJUEIRO, Daniel O.; CELESTINO, Victor R. R. A *Comprehensive Review of Artificial Intelligence Regulation:* Weighing Ethical Principles and Innovation. 09.09.2024. Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4950727.

> *trained to imitate different personalities, adversarial parties can flood social media with fake speech. The concern is not necessarily that these bot accounts will all be effective, but rather that they will engender a sense of general distrust among the population.*[33]

This is particularly true in the case of deepfakes.[34] According to a report by the International Panel on the Information Environment (IPIE), 80% of the countries holding elections in 2024 had incidents with generative AI and 90% of the incidents involved content creation (audio, messages, images, videos and social media posts). Moreover, 69% of the incidents were described as having a harmful role in the election.[35] The increasing sophistication of AI-generated content – such as highly realistic videos or audio recordings – and the scalability of these tools reinforce the technology's disruptive potential.

In particular, generative AI "*may be used to derive numerous personalized versions of political messages from a single human-designed template*",[36] making it possible to tailor content to different personality types in an automated and scalable manner. In the context of political microtargeting (or data-driven targeted advertising),[37] this

---

[33] ARBEL, Yonathan A.; TOKSON, Matthew J.; LIN, Albert. Systemic Regulation of Artificial Intelligence (December 16, 2023). *AI Safety Legal Paper Series* 12-24, 56 Ariz. St. L. J. 545, University of Utah College of Law Research Paper No. 582, Available at SSRN: https://ssrn.com/abstract=4666854, p. 569-570.

[34] Deepfakes can be defined as "a set of techniques used to synthesise new visual products, for example by replacing faces in the originals". FLORIDI, Luciano. Artificial intelligence, deepfakes and a future of ectypes. *Philosophy and Technology*. 2018;31:317–321. doi: 10.1007/s13347-018-0325-3

[35] International Panel on the Information Environment [I. Trauthig, P. N. Howard, S. Valenzuela (eds.)]. *The Role of Generative AI Use in 2024 Elections Worldwide*. Zurich, Switzerland: IPIE, 2025. Technical Paper, TP2025.2, doi: 10.61452/HZUE9853.

[36] Simchon, Almog; Edwards, Matthew; Lewandowsky, Stephan. The persuasive effects of political microtargeting in the age of generative artificial intelligence, *PNAS Nexus*, Volume 3, Issue 2, February 2024, pgae035, https://doi.org/10.1093/pnasnexus/pgae035

[37] There is still no shared definition of political microtargeting, but, in general, it consists of "the collection and use of personal data in order to formulate and send political messages to targets and exert political influence". KRÖGER, Jacob Leon; ERRENST, Emilia; NAU, Niklas; OJANPERÄ, Sanna, What is Political Microtargeting (May 22, 2024). This chapter is an excerpt from: Kröger, J., Errenst, E., Nau, N., and Ojanperä, S. (2024). *Mitigating the Risks of Political*

technology can "*take any political message and derive versions that target people of different personalities in an automated manner and at scale*".[38] he convergence of generative AI and social media platforms thus enables the dissemination of malicious content with a spread, scale, and sophistication that are simply incomparable to traditional forms of political communication.

Real-world examples of AI misuse in electoral contexts have already been reported in several countries. In the United States, robocalls that used the voice of President Joe Biden discouraged people from voting in the New Hampshire primaries in 2024.[39] Also, in August 2024, Trump posted AI images of Taylor Swift and her fans endorsing his campaign for president,[40] and, in September 2024, a fake video accused the candidate Kamala Harris of being involved in a hit-and-run accident, and was seen on social media millions of times.[41] In Slovakia, a viral fake audio that circulated on Facebook showed the leader of the liberal progressive Slovakia party discussing how to rig the parliamentary election.[42] During the European Parliament elections in

Microtargeting – Guidance for Policymakers, Civil Society, and Development Cooperation, pp. 7-9. (forthcoming), Available at SSRN: https://ssrn.com/abstract=4856178 or http://dx.doi.org/10.2139/ssrn.4856178

38 Simchon, Almog; Edwards, Matthew; Lewandowsky, Stephan. The persuasive effects of political microtargeting in the age of generative artificial intelligence, *PNAS Nexus*, Volume 3, Issue 2, February 2024, pgae035, https://doi.org/10.1093/pnasnexus/pgae035

39 The incident even led to a $6 million fine imposed on political consultant Steve Kramer, who was responsible for creating the deepfake audio by the FCC - Federal Communications Commission (the agency that regulates telecommunications and broadcasting in the United States). DEL VALLE, Gaby. Political Consultant Behind the Joe Biden Deepfake Robocalls Faces $6 Million Fine. *The Verge*, May 23, 2024. Available at: https://www.theverge.com/2024/5/23/24163411/fcc-fine-biden-deepfake-robocalls-steve-kramer-lingo-telecom.

40 ROBINS-EARLY, Nick. How did Donald Trump end up posting Taylor Swift deepfakes? The Guardian, 26.08.2024. Available at: https://www.theguardian.com/technology/article/2024/aug/24/trump-taylor-swift-deepfakes-ai

41 The fake video was linked back to a Russian disinformation campaign. See: MENN, Joseph. Russians made video falsely accusing Harris of hit-and-run, Microsoft says. The Washington Post, 17.09.2024. Available at: https://www.washingtonpost.com/technology/2024/09/17/russia-fake-videos-election-interference-kamala-harris/

42 DEVINE, Curt; O'SULLIVAN, Donie; LYNGAAS, Sean. A fake recording of a candidate saying he'd rigged the election went viral. Experts say it's only the beginning. *CNN*, 01.02.2024. Available

June, 2024, a deep fake video created by left-wing activists in Germany showed Chancellor Olaf Scholz calling for an Alternative für Deutschland (AfD) party ban.[43] In Bangladesh, a fake video announced a Bangladeshi political candidate's withdrawal from elections.[44] In South Africa, a deepfake video showed U.S. rapper Eminem endorsing the Economic Freedom Fighters party and criticizing the African National Congress.[45] In Argentina, during the 2023 electoral campaign, candidates Javier Milei and Sergio Massa widely shared AI-generated images and videos in their campaigns, both for self-promotion and to attack their opponent.[46]

The most notable example occurred in Romania's presidential election in 2024.[47] In the first (preliminary) round of voting, Călin Georgescu, a relatively unknown far-right candidate, unexpectedly took the lead, reportedly benefiting from non-transparent campaign financing, the use of digital influencers, and TikTok's recommendation algorithm, which significantly amplified his visibility. The online environment was flooded with short-form videos, conspiracy content, and deepfakes, contributing to voter confusion. TikTok failed to detect or flag these materials as campaign content or to take action against coordinated disinformation efforts.[48] The first round

at: <https://edition.cnn.com/2024/02/01/politics/election-deepfake-threats-invs/index.html>.

43 European Partnership for Democracy (EPD). *Artificial Intelligence and Elections: Protecting Electoral Integrity in the Age of AI*. September 2024. Available at: https://epd.eu/content/uploads/2024/09/AI-and-elections.pdf.

44 ELLIOTT, Vittoria. The WIRED AI Elections Project. *Wired*, 30.05.2024. Available at: <https://www.wired.com/story/generative-ai-global-elections/>.

45 COSSER, Kirsten. Will the real Slim Shady please stand up? Eminem video endorsing South African opposition party EFF and bashing the ruling ANC is a deepfake. Africa Check, 05.03.2024. Available at: <https://africacheck.org/fact-checks/meta-programme-fact-checks/will-real-slim-shady-please-stand-eminem-video-endorsing>.

46 NICAS, Jack; HERRERA, Lucía Cholakian. Las campañas electorales de Argentina recurren a la IA. *The New York Times*, 15.11.2023. Available at: <https://www.nytimes.com/es/2023/11/15/espanol/elecciones-argentina-imagenes-inteligencia-artificial.html>.

47 MYERS, Steven Lee; THOMPSON, Stuart A. A.I. Is Starting to Wear Down Democracy. *The New York Times*, June 26, 2025. Available at: <https://www.nytimes.com/2025/06/26/technology/ai-elections-democracy.html>.

48 RADU, Roxana. Expert Comment: Which urgent tech lessons must the EU take from Romania's election? Oxford News, 22 may 2025. Available at: <https://www.ox.ac.uk/news/2025-05-22-expert-comment-which-urgent-tech-lessons-must-eu-take-romania-s-election>.

of Presidential election was then annulled by the Romanian Constitutional Court.[49] As stated by a report by The New York Times, " *It was the first major election in which A.I. played a decisive role in the outcome. It is unlikely to be the last*".[50]

This global trend underscores the urgent need to establish legal and institutional safeguards to regulate the use of AI in elections. In response, international organizations such as the United Nations (UN) have advocated for a comprehensive and collaborative approach to AI governance, calling for the definition of global standards and ethical guidelines to mitigate long-term risks (horizontal approach).[51] Additionally, major technology companies such as Google, Meta, OpenAI, X (formerly Twitter), and TikTok have entered into voluntary agreements to curb the deceptive use of AI in the 2024 elections.[52] States are also discussing and adopting new rules to regulate the use of AI in elections, through different approaches (such as imposing fines and political sanctions, as also banning the use of deepfakes in political campaigns etc).[53]

---

[49] About the decision: CARROZZINI, Alina. Shooting Democracy in the Foot? The Romanian Constitutional Court's Annulment of Presidential Elections. Verfassungsblog, 13 December 2024. Available at: <https://verfassungsblog.de/shooting-democracy-in-the-foot/>

[50] MYERS, Steven Lee; THOMPSON, Stuart A. A.I. Is Starting to Wear Down Democracy. *The New York Times*, June 26, 2025. Available at: <https://www.nytimes.com/2025/06/26/technology/ai-elections-democracy.html>.

[51] United Nations. Can artificial intelligence (AI) influence elections? United Nations, 07.06.2024. Available at: <https://unric.org/en/can-artificial-intelligence-ai-influence-elections/>

[52] MSC. A Tech Accord to Combat Deceptive Use of AI in 2024 Elections. MSC. Available at: <https://securityconference.org/en/aielectionsaccord/>. Accessed on: 12.08.2024. Noticeable that the accord has many weaknesses: "the commitments are both somewhat vague and entirely voluntary. There is no mechanism in the accord by which companies are required to report on their progress. Nor are there any suggested benchmarks for others to track their progress." NORDEN, Lawrence. HARRIS, David Evan. New Tech Accord to Fight AI Threats to 2024 Lacks Accountability for Companies. Brennan Center for Justice, 29.02.2024. Available at: https://www.brennancenter.org/our-work/analysis-opinion/new-tech-accord-fight-ai-threats-2024-lacks-accountability-companies

[53] "South Korea amended its Public Official Election Act, banning the use of AI-generated deepfakes in political campaigns for 90 days leading up to election day, with exceptions for deepfakes that promote voter participation and intraparty activities (e.g., AI-generated campaign mottos, song lyrics or speeches). Brazilian Superior Electoral Court has also approved 12 resolutions that

The increasing influence of AI and social media platforms on elections underscores the need for robust regulatory frameworks to safeguard democratic processes globally. This requires not only a general (horizontal) approach, based on principles applicable across all sectors, but also a sector-specific (vertical) strategy tailored to the unique challenges posed by the use of AI in elections. In this context, Brazil has taken proactive steps to address these challenges, particularly through TSE. The next section examines one such initiative: Resolution nº 23.732/2024 from the TSE, which aims to regulate AI use in electoral campaigns and mitigate disinformation risks.

3. THE BRAZILIAN RESPONSE: RESOLUTION Nº 23.732/2024 FROM THE SUPERIOR ELECTORAL COURT

3.1 CONTEXTUAL BACKGROUND: ELECTORAL INSTITUTIONS AND DISINFORMATION CHALLENGES

The Brazilian electoral system relies on the role of tribunals in its supervision – much in the sense of the Argentinean model in which the *Cámara Nacional Electoral* is part of the judicial branch.[54] The Brazilian electoral courts date back from the 1930s. The 1988 Constitution maintained the system in which electoral courts administrate, rule upon, and regulate elections.

The Superior Electoral Court (TSE), the highest structure within the Electoral Justice system, plays a fundamental role in the construction and employment of

will guide the upcoming elections. One of the resolutions providing for electoral propaganda prohibits the use of deepfakes, restricts the use of chatbots and avatars to simulate dialogue between a candidate and a real person and requires disclosure for AI-generated and AI-manipulated content. In the US, several states have implemented bills that safeguard against the deceptive use of AI in elections. Other efforts to address the threats of AI use in elections include the European Union's Artificial Intelligence Act (EU AIA) and the UK's response paper to its white paper on AI" SOON, Carol. QUEK, Samantha. Safeguarding elections from threats posed by artificial intelligence. IPS Working Papers No. 56, August, 2024. Available at: https://lkyspp.nus.edu.sg/docs/default-source/ips/ips-working-paper-no-56_safeguarding-elections-from-threats-posed-by-artificial-intelligence.pdf

[54] https://www.electoral.gob.ar/nuevo/paginas/cne/competencia_cne.php

Brazilian democracy. Its main powers are established by the 1988 Constitution and the Electoral Code and they encompass jurisdictional, administrative and normative competences. The normative role of the TSE is described in Article 1, sole paragraph, and Article 23, IX, of the Electoral Code. This function allows the Court – through resolutions – to issue instructions for the enforcement of electoral laws, including the Electoral Code itself. The content of these regulations aims to govern matters within the jurisdiction of the collegiate body that established them, creating general and abstract provisions. Within this framework, the TSE exercises both normative and administrative powers, including the issuance of binding resolutions that apply to all levels of government.

This institutional design grants the TSE a central role in preserving the legitimacy and integrity of the democratic process. As digital technologies have become increasingly embedded in electoral dynamics – particularly through the use of social media and generative AI, as discussed in section 2 – the Court has been called upon to respond to new forms of interference, manipulation, and disinformation that threaten electoral fairness.

These challenges came to the forefront during the 2018 presidential election, widely regarded as a disruptive moment in Brazilian democracy. For the first time in the country's democratic history, a candidate who did not meet the traditional prerequisites for electoral success was elected, amid a campaign marked by an unprecedented volume of digital disinformation.[55] This experience highlighted the urgent

[55] On the 2018 elections: NICOLAU, Jairo. *O Brasil dobrou à direita:* uma radiografia da eleição de Bolsonaro em 2018. Rio de Janeiro: Zahar, 2020 [e-book], ch. 1. RUEDIGER, Marco Aurélio. GRASSI, Amaro FGV DAPP. *Redes sociais nas eleições 2018*: Policy Paper 1. Rio de Janeiro: FGV DAPP, 2018, p. 15-22. MACHADO, Caio; KIRA, Beatriz; HIRSCH, Gustavo; MARCHAL, Nahema; KOLLANYI, Bence; HOWARD, Philip N.; LEDERER, Thomas; BARASH, Vlad. News and Political Information Consumption in Brazil: Mapping the First Round of the 2018 Brazilian Presidential Election on Twitter. Data Memo 2018.4. Oxford, UK: *Project on Computational Propaganda*. CARVALHO, Lucas Borges de. A democracia frustrada: fake news, política e liberdade de expressão nas redes sociais. *Internet & Sociedade*, V.1, N.1 – 2020. SANTOS, João Guilherme dos. FREITAS, Miguel. ALDÉ, Alessandra. SANTOS, Karina. CUNHA, Vanessa. WhatsApp, política mobile e desinformação: a hidra nas eleições presidenciais de 2018. *Comunicação & Sociedade*. V. 41, n. 2, 2019, p. 307-334. MARQUES, Ana Luiza Pinto Coelho. *Erosão constitucional na era digital*: uma análise a partir do processo eleitoral brasileiro de 2018. São Paulo: Editora

need to strengthen regulatory frameworks for digital governance and raised concerns about the impact of private platform decisions on core democratic values.

In response, the TSE launched a series of institutional initiatives to address these emerging threats in elections. For example, the TSE has implemented the Electoral Justice Permanent Program on Countering Disinformation, which seeks to reduce the harmful effects of disinformation regarding the electoral court and its members, the electronic voting system, and the electoral process.[56] Among the measures adopted in the program, TSE has established partnerships with social media platforms, which include specific measures to address the problem of misinformation in the electoral process – although recent developments in the 2024 U.S. presidential election suggest that major tech companies may already exhibit political bias.[57] These measures include the provision of a content label that directs users to official information about the election and the creation of an extrajudicial communication channel dedicated to the TSE for reporting content that spreads disinformation related to elections.[58] Another example is Resolution nº 23.714, published in 2022, which focused on tackling disinformation that threatens the integrity of the electoral process, by mandating the removal of misleading or false content related to elections from digital platforms and laying out procedures for the immediate removal of

Dialética, 2024. MEYER, Emilio Peluso Neder. *Constitutional Erosion in Brazil*. Oxford; New York: Hart Publishing, 2021.

56 Regarding the actions already imposed so far, the highlight is the signing of partnerships with digital platforms, with the agreement on measures such as: signaling electoral content and directing to official information about the electoral process; the joint development of a new chatbot (virtual assistant) for interacting with participants; creation of search notices to assist in the search for information about the electoral process; and assistance in disseminating content produced by the official TSE contact. Brazilian Superior Electoral Court (TSE). *Countering Disinformation Program for the 2022 Elections*. Available at: https://international.tse.jus.br/en/misinformation-and-fake-news/tse-brazil-counter-disinformation-program-2022-f.pdf. Accessed on: 05.10.2024.

57 WENDLING, Mike. "Trump sides with tech bosses in Maga fight over immigrant visas", *BBC News*, 24 Dec. 2024, https://www.bbc.com/news/articles/clyv7gxp02yo.

58 Brazilian Superior Electoral Court (TSE). Eleições 2022: TSE e plataformas digitais firmam parceria para combate à desinformação. TSE, 07.02.2022. Available at: https://www.tse.jus.br/comunicacao/noticias/2022/Fevereiro/eleicoes-2022-tse-e-plataformas-digitais-firmam-parceria-para-combate-a-desinformacao

harmful disinformation that could affect the fairness of elections.[59]

Despite these efforts, the 2022 general elections were also marked by the increased use of technology for political purposes. This included the rise of content criticizing and attacking democratic and constitutional institutions such as the TSE, the Federal Supreme Court (STF), and the electronic ballot box,[60] as well as the use of artificial intelligence to produce false or misleading information.[61] This process was aggravated by Jair Bolsonaro's refusal to explicitly recognize the victory of Luiz Inácio Lula da Silva in the presidential race and the fairness of the electoral process, culminating in the invasion of the buildings of three branches of power in Brasília by Bolsonaro supporters on January 8, 2023, who claimed the electoral process was rigged.[62]

---

[59] Brazilian Superior Electoral Court (TSE). *Resolution No. 23.714 of October 20, 2022*. Available at: https://www.tse.jus.br/legislacao/compilada/res/2022/resolucao-no-23-714-de-20-de-outubro-de-2022.

[60] In 2018, the focus of the disinformation campaign was on anti-communist content, the defense of religious values of the so-called traditional family, and content against the Brazilian Worker's Party (*Partido dos Trabalhadores*), the party of the incumbent president, Luis Inacio Lula da Silva. However, this pattern changed in subsequent elections. In the 2020 municipal elections and the 2022 presidential election, there was an increase in content criticizing and attacking democratic and constitutional institutions such as the TSE, the Federal Supreme Court (STF), and the electronic ballot box. As noted in a report by FGV DAPP: "there is a discursive difference between narratives of electoral distrust from 2018, which were characterized by claims of collusion and institutional capture by left-wing parties, particularly the PT (…) the debate has moved somewhat away from narratives tied to communism, maintaining a conspiratorial tone but focusing more on issues of transparency and security" (author's translation). RUEDIGER, Marco Aurelio. GRASSI, Amaro (coord.). O ecossistema digital nas eleições municipais de 2020 no Brasil: o buzz da desconfiança no sistema eleitoral no Facebook, Youtube e Twitter. Policy Paper. Rio de Janeiro: FGV DAPP, 2020, p. 22

[61] In the 2022 presidential elections, there were already cases of artificial intelligence being used to produce false content involving the main presidential candidates. For instance, a video circulated in which journalist Renata Vasconcellos appeared to announce a false poll result (available at https://www.youtube.com/watch?v=X3tU01FDPhA, accessed on 08.08.2024). Additionally, the Supreme Federal Court ordered the removal of a deepfake video that showed then-presidential candidate Luiz Inácio Lula da Silva being booed at an event in Uberlândia, Minas Gerais (available at https://www.cnnbrasil.com.br/politica/carmen-lucia-manda-remover-videos-contra-lula-com-deep-fake/, accessed on 08.08.2024)

[62] MEYER, Emilio Peluso Neder. BUSTAMANTE, Thomas. Accountability in Brazil: Between

That is the reason why the TSE kept on implementing a series of measures aimed at confronting misinformation and fake news, as well as reinforcing the security, auditability, and transparency of the voting system and the electronic ballot boxes. The goal is to enhance the credibility of Brazilian Electoral Justice. These measures are aligned with TSE's constitutional mission to ensure that the electoral process takes place legitimately and democratically, which fall within the Court's administrative and normative competence.

3.2 RESOLUTION Nº 23.732/2024: REGULATORY INNOVATIONS AND ENFORCEMENT MECHANISMS

In attention to the rise of generative AI and its potential impacts in elections, the TSE introduced a new regulation for the use of artificial intelligence in electoral campaigns and political communication on social media platforms with which candidates, parties, and application providers must comply with. Resolution nº 23.732, among other things, establishes the absolute prohibition of deepfake use in electoral advertising, the requirement of identification labels for synthetic multimedia content generated by AI, the restriction of the use of chatbots and avatars to mediate campaign communication, and the prohibition of promoting content that supports negative propaganda or that spreads false data, fraudulent news, or blatantly untrue or severely decontextualized facts, even if beneficial to the user responsible for promoting it.

To ensure compliance with these rules, the resolution, aligned with TSE's administrative competence and to its police authority,[63] states that electoral judges may order the immediate removal of content from the internet that, in its form or means of dissemination, violates its provisions. As a rule, when the irregularity is concerned with the content of the advertisement, police power cannot be exercised. In this case,

Democracy and the Digital Age." *Verfassungsblog*, May 8, 2023. https://verfassungsblog.de/accountability-in-brazil/.

[63] The police power over electoral propaganda is not new in Brazil. It was established by Law nº 9.504/1997, to prevent illegal practices that could harm the electoral process. Nevertheless, prior censorship of the content of programs to be broadcast on television, radio, or the internet is prohibited.

any irregularity must be reported to the Electoral Public Prosecutor's Office.

However, Resolution nº 23.732 specifies that, in cases where electoral propaganda on the internet disseminates notoriously false or severely out-of-context information about the electronic voting system, the electoral process, or the Electoral Justice, judges are bound to the collective decisions of the Superior Electoral Court on the same matter. This means that when the TSE has already ruled on the removal or maintenance of identical content, judges, acting within their police powers, may issue removal orders based on these previous TSE decisions. To facilitate this process, the Resolution provides the creation of a public repository, which will contain decisions from the TSE regarding content removal. After receiving a removal order, applications providers must report compliance via the system and update the repository, when instructed to do so.

Regarding politicians, non-compliance with the rules of the new resolution may result in the revocation of their registration or mandate, including and expressly in cases where the obligation to not use any form of "fabricated or manipulated content to disseminate notoriously untrue or decontextualized facts with the potential to cause harm to the balance of the election or the integrity of the electoral process" is violated. This provision is in line with the rulings made by the TSE in the judgment of relevant cases, such as that of the state representative from the state of Paraná, Eduardo Fernando Francischini – who got his mandate revoked and was declared ineligible – and the case that culminated in the declaration of ineligibility of Jair Bolsonaro for a period of eight-years. In both cases, the Court has decided that the dissemination of false information about the electronic voting system can lead to revocation of a mandate and a ruling of ineligibility.[64]

Regarding application service providers that allow the dissemination of political-electoral content, the Resolution requires them to adopt measures to prevent or reduce the circulation of notoriously untrue or seriously out-of-context facts that

[64] Brazilian Superior Electoral Court (TSE). Ordinary Electoral Appeal 060397598/PR, Relator: Min. Luis Felipe Salomão, Judgment: 28/10/2021, Published in the Electronic Justice Gazette-228, date 10/12/2021. Brazilian Superior Electoral Court (TSE). Action for Electoral Judicial Investigation 060081485/DF, Relator: Min. Benedito Gonçalves, Judgment: 30/06/2023, Published in the Electronic Justice Gazette-147, date 02/08/2023

could affect the integrity of the elections. Among these measures are the implementation of accessible notification tools and reporting channels, the improvement of content recommendation systems, the assessment of the impact of their services on the integrity of the electoral process, and the enhancement of their technological and operational capabilities. Additionally, there is an obligation to maintain a repository of advertisements (for application providers that offer services for boosting political-electoral content) and to provide an accessible and user-friendly query tool for advanced searches in the repository data. Furthermore, Resolution nº 23.732 prohibits the paid boosting of content that contains notoriously false and severely decontextualized facts. It also requires that, when illegal content is detected or when users notify its circulation, providers must, immediately, stop the promotion, monetization, and access to the content.[65] If they do not comply with this duty during the electoral period, providers could be jointly held liable under civil and administrative law.[66]

In order to enhance compliance with Resolution nº 23.732/2024, TSE signed agreements with digital platforms, including Google, LinkedIn, X, Kwai, TikTok, Telegram, Facebook, WhatsApp, and Instagram. The agreements foster a collaborative approach between public authorities and digital platforms to tackle the challenges posed by disinformation and AI. Among the key innovations, the platforms have committed to working alongside the Integrated Center for Combating Disinfor-

---

65 The swift removal of potentially harmful information is essential for the protection of the electoral process, especially considering the rapid spread of false content and the damaging potential that arises from extended exposure to false or misleading information. Studies demonstrate the high capacity for the spread of false news and content that incites violence compared to true and positive content. VOSOUGHI, Soroush; ROY, Deb; ARAL, Sinan. The spread of true and false news online. Science, VOL. 359, NO. 6380, p. 1,146. MATHEW, Binny; DUTT, Ritam; GOYAL, Pawan; MUKHERJEE, Animesh. Spread of hate speech in online social media. ArXiv:1812.01693, 2019. Available at: https://arxiv.org/abs/1812.01693.

66 In this point, the Resolution was controversial in Brazil, since, up to this moment, there is no specific law exclusively addressing social media platforms or governing content moderation activities. However, Article 19 of Law nº 12.965/2014 — known as the Brazilian Civil Rights Framework for the Internet (Marco Civil da Internet - MCI) — conditions the liability of application providers for hosting illicit content generated by third parties on the requirement of a specific court order to remove the content. The constitutionality of this legal provision is currently under review by the Supreme Federal Court.

mation and Defending Democracy (*Centro Integrado de Enfrentamento à Desinformação e Defesa da Democracia – CIEDDE*), which coordinates efforts to fight disinformation. Additionally, the Electoral Disinformation Alert System (*Sistema de Alertas de Desinformação Eleitoral – SIADE*) has been introduced, enabling citizens to report suspicious content. Importantly, the companies have reaffirmed their commitment to swiftly removing or limiting the spread of content deemed to be disinformation.

Therefore, Resolution nº 23.732/2024 represents a comprehensive effort by the Brazilian Superior Electoral Court to regulate the use of AI in electoral campaigns, strengthen measures against disinformation, and ensure the integrity of the electoral process, including specific requirements for content removal, penalties for non-compliance, and collaboration with digital platforms to combat misleading information.

3.3 COMPARATIVE PERSPECTIVES: INTERNATIONAL APPROACHES TO AI AND ELECTORAL INTEGRITY

The TSE Resolution stands out internationally by imposing comprehensive measures to address the risks posed by AI and disinformation in elections. As seen in section 3.2., it represents an example of sectorial regulation of AI, issued by the institution responsible for supervising and regulating elections in the country. By focusing on the specific use of AI in political campaigns, the Resolution illustrates how targeted rules can respond more effectively to context-specific challenges. However, in the absence of a general AI law in Brazil, this sectorial initiative operates in a "legal vacuum", with no overarching normative framework regarding AI to align with or complement its provisions.

Although Brazil has adopted a comprehensive regulatory response through Resolution nº 23.732/2024, it is not the only attempt to regulate the use of AI in elections. However, international examples of specific regulations on the use of AI in elections remain scarce, [67] and many of the existing initiatives are narrower in scope compared

[67] See: SOON, Carol. QUEK, Samantha. Safeguarding elections from threats posed by artificial intelligence. IPS Working Papers No. 56 (August 2024). Available at:

to the Brazilian Resolution — often focusing solely on the prohibition of deepfakes, without broader provisions or enforcement mechanisms directed at social media platforms.

Notable examples include South Korea's amendment to its Public Official Election Act, which took effect in January 2024. Under this provision, from 90 days before the election until election day, it is prohibited to produce, edit, distribute, screen, or post virtual audio, images, or videos that are difficult to distinguish from reality and created using artificial intelligence or similar technologies for the purpose of election campaigning. Violations are punishable with imprisonment (up to seven years) or by a fine. Outside of the 90-day pre-election period, such content may be used in campaigning, but it must include a clear notice indicating that the material is virtual and was generated using artificial intelligence or other technologies. In this case, violations are punishable with a fine.[68]

In Singapore, the Elections (Integrity of Online Advertising) Bill was amended to prohibit the publication, boosting, sharing and reposting of election-related online advertisements containing audio, visual, or audiovisual realistic, but false, representations of a candidate, if the representation is realistic enough that members of the public could reasonably believe it to be true.[69] The provision targets content generated or manipulated, at least in part, using digital technologies such as AI. During the

---

https://lkyspp.nus.edu.sg/docs/default-source/ips/ips-working-paper-no-56_safeguarding-elections-from-threats-posed-by-artificial-intelligence.pdf. SCHULDT, Lasse. Every Fake You Make: Blanket Deepfake Bans Are the Next Level in Asia's War on Fake News. Verfassungsblog, October 9, 2024. Available at: https://verfassungsblog.de/fake-news-content-moderation-singapoer-south-korea-election/.

[68] SCHULDT, Lasse. Every Fake You Make: Blanket Deepfake Bans Are the Next Level in Asia's War on Fake News. *Verfassungsblog,* 09 October 2024. Available at: <https://verfassungsblog.de/fake-news-content-moderation-singapoer-south-korea-election/>.

[69] https://sso.agc.gov.sg/Bills-Supp/29-2024/Published/20240909?DocDate=20240909. See also: TAN, Bryan; KONG, Hannah; GOH, Eng Han. Singapore bans deepfakes in elections. ReedSmith, 04 November 2024. Avilable at: <https://www.reedsmith.com/en/perspectives/2024/11/singapore-bans-deepfakes-in-elections>. SCHULDT, Lasse. Every Fake You Make: Blanket Deepfake Bans Are the Next Level in Asia's War on Fake News. *Verfassungsblog,* 09 October 2024. Available at: <https://verfassungsblog.de/fake-news-content-moderation-singapoer-south-korea-election/>.

election period, violations may result in fines, imprisonment, or both.

In the United States, state legislatures across the country are regulating deepfakes in elections.[70] Rather than prohibiting the use of synthetic or altered content outright, usually, state laws addressing deepfakes require certain safeguards.[71] Most states either mandate disclaimers indicating that the content has been digitally manipulated or restrict the timing of its use – often allowing it only outside of a defined period leading up to Election Day. In some cases, both requirements apply simultaneously. Penalties for violating these laws vary by state, but most include injunctive relief – such as court-ordered removal of manipulated content – as the primary enforcement mechanism, alongside criminal and civil sanctions.[72]

Some jurisdictions do not adopt sector-specific regulation for AI in elections but apply general AI rules that extend to electoral contexts. This is the case of China, where the *Provisions on the Administration of Deep Synthesis Internet Information Services* establish general obligations for providers and users of synthetic media. Among other rules, Article 6 prohibits the use of deep synthesis technologies to produce or disseminate fake news information. Articles 7, 10 and 11 require providers to ensure information security, report illegal content, store relevant records, take action against users, and implement mechanisms to dispel false information. Article 17 mandates the labeling of synthetic content that may mislead or confuse the public.[73]

---

[70] Public Citizen. Tracker: State Legislation on Deepfakes in Elections. Public Citizen. Available at: <https://www.citizen.org/article/tracker-legislation-on-deepfakes-in-elections/>.

[71] Idaho formally prohibits the use of manipulated media in campaign materials. However, the law "provides for an "affirmative defense for any action brought…[if] the electioneering communication containing synthetic media includes a disclosure stating, "This (video/audio) has been manipulated...". LARKIN, CJ. Regulating Election Deepfakes: A Comparison of State Laws. Tech Policy Press, 8 January 2025. Available at: <https://www.techpolicy.press/regulating-election-deepfakes-a-comparison-of-state-laws/>.

[72] LARKIN, CJ. Regulating Election Deepfakes: A Comparison of State Laws. Tech Policy Press, 8 January 2025. Available at: <https://www.techpolicy.press/regulating-election-deepfakes-a-comparison-of-state-laws/>.

[73] China Law Translate. Provisions on the Administration of Deep Synthesis Internet Information Services. China Law Translate, 11.12.2022. Available at: <https://www.chinalawtranslate.com/en/deep-synthesis/>.

As noted in section 1, the European Union adopts a general regulatory approach to AI.[74] The AI Act, along with other frameworks – such as the Digital Services Act (DSA), the General Data Protection Regulation (GDPR), and the Digital Markets Act (DMA) – applies to electoral contexts. The AI Act explicitly references the protection of democracy, the rule of law, and fundamental rights, including the right to vote and to stand as a candidate in European Parliament elections. Recital 48 of the AI Act also identifies as high-risk those AI systems that may negatively impact fundamental rights, including the right to vote. However, the Act does not clearly list electoral-related AI applications among the high-risk categories, leaving potential gaps in addressing risks to electoral integrity.[75]

The comparative overview reveals that, while other jurisdictions have taken steps to address the risks of AI in electoral contexts, the Brazilian approach stands out for its institutional centralization, regulatory breadth, and enforcement mechanisms. Unlike other examples of sectorial regulation – such as those adopted in South Korea, Singapore, and by U.S. state legislatures – which typically focus on banning or restricting deepfakes, imposing labeling requirements, and applying penalties (such as fines or imprisonment), Resolution nº 23.732/2024 articulates a more comprehensive and enforceable framework. Importantly, the Resolution is not merely repressive; it also adopts a proactive stance. Several of its provisions aim to foster an institutional infrastructure and regulatory culture that upholds electoral integrity before harms occur, notably by establishing preventive obligations for application service providers that enable the dissemination of political-electoral content. Also, by assigning obligations not only to political actors but also to platforms, and by providing clear sanctions for noncompliance, the Brazilian regulation addresses both the production and the circulation of electoral misleading content.

With regard to jurisdictions that do not have specific norms to regulate AI in elections, the Brazilian regulation stands out for being explicitly tailored to the dynamics of the electoral process. In the European Union, for instance, the AI Act –

---

74 https://www.europarl.europa.eu/doceo/document/TA-9-2024-0138_EN.pdf

75 RABITSCH, Armin; CALABRESE, Sofia. The EU's Artificial Intelligence Act and its Impact on Electoral Processes: a Human Rights-Based Approach. European Partnership for Democracy, September 2024. Available at: < https://epd.eu/content/uploads/2024/09/AI-and-elections.pdf>.

despite its normative ambition – still lacks detailed attention to electoral applications of AI. It does not explicitly address the risks posed to elections nor does it establish mitigation measures targeted specifically at those risks.[76] In China, the *Provisions on the Administration of Deep Synthesis Internet Information Services* impose general obligations on providers and users of synthetic media, encompassing both repressive and proactive duties. However, despite their broader scope, these rules do not include provisions directly applicable to elections, such as obligations for political candidates or explicit political sanctions for violations. This divergence highlights the importance of sector-specific and context-sensitive regulation, particularly in high-risk domains such as elections, where the misuse of AI can have immediate and far-reaching consequences for electoral trust and democratic stability.

At the same time, the Brazilian model also reveals important limitations. In the absence of a general AI law, Resolution nº 23.732/2024 operates in a fragmented regulatory environment, lacking overarching principles or safeguards that would ensure consistency across different sectors and uses of AI. While the Resolution imposes specific duties on application providers in the electoral context, in the absence of a general law, Brazil still lacks broader structural obligations – such as transparency in content moderation systems, auditability, explainability, access to data or prohibition/restrictions on the use of sensitive personal information to content personalization – all of which are essential to address systemic risks posed by recommendation systems and political microtargeting. Therefore, while Brazil offers a regulatory model in the electoral sector, its experience also highlights the need for integrated frameworks that combine sectorial depth with systemic consistency.

Moreover, the effectiveness of the Resolution will depend largely on its judicial interpretation and enforcement. As it assigns a central role to electoral judges in identifying and ordering the removal of harmful content, the development of a robust and coherent jurisprudence will be key to ensuring clarity regarding the scope of the rules and the legal consequences of non-compliance. In the absence of such clarity, the Resolution may fall short of achieving its intended impact on electoral

[76] RABITSCH, Armin; CALABRESE, Sofia. The EU's Artificial Intelligence Act and its Impact on Electoral Processes: a Human Rights-Based Approach. European Partnership for Democracy, September 2024. Available at: < https://epd.eu/content/uploads/2024/09/AI-and-elections.pdf>.

integrity and democratic resilience. The judicialization of electoral content moderation can also generate tensions between freedom of expression and electoral fairness, requiring careful balancing by courts and public authorities.

In practical terms, the Resolution also faces challenges related to institutional capacity, technical expertise, and platform cooperation. Despite formal agreements with major technology companies, questions remain regarding their willingness and ability to comply with electoral court decisions, especially in a fast-evolving digital ecosystem. Evidence from Brazil's municipal elections in October 2024 illustrates these difficulties: electoral deepfakes circulated widely across different social media platforms, primarily targeting mayoral candidates. The content included: (i) deepfake pornography aimed at female candidates; (ii) impersonations of media figures, using the faces of well-known Brazilian journalists in AI-manipulated videos; (iii) deepfakes intended to mislead voters about political events, ballots, voting locations, or other critical electoral information; (iv) the unauthorized use of images of prominent Brazilian politicians to support or undermine candidates; and (v) humorous deepfakes.[77] So far, the Superior Electoral Court (TSE) has not issued rulings on the political consequences for candidates or elected officials who shared AI-generated content in the 2024 municipal elections; only regional electoral courts have ruled on these cases to date.[78]

CONCLUSION

As generative AI tools become increasingly accessible, their use in electoral processes presents one of the most pressing challenges for contemporary democracies. The risks posed by synthetic content – ranging from deepfakes to microtargeted disinformation – require specific legal frameworks, able to address the particular

---

[77] FARRUGIA, Beatriz. Brazil's electoral deepfake law tested as AI-generated content targeted local elections. *Digital Forensic Research Lab* (DFRLab), November 26, 2024. Available at: <https://dfrlab.org/2024/11/26/brazil-election-ai-deepfakes/>.

[78] On the cases judged by regional electoral courts, see: GUIMARÃES, Nino. Justiça Eleitoral determina remoção de deepfakes com base na qualidade da manipulação. Jota, 26 October 2024. Available at: < https://www.jota.info/eleicoes/justica-eleitoral-determina-remocao-de-deepfakes-com-base-na-qualidade-da-manipulacao>

risks that arise from the use of AI in a high-risk sector such as elections. These legal frameworks must not only be reactive but also proactive, capable of adapting to evolving technologies and safeguarding the integrity of democratic deliberation.

In this context, Brazil's Resolution nº 23.732/2024 stands out as a significant regulatory innovation. Adopted by the Superior Electoral Court (TSE), the Resolution represents a pioneering example of sectorial AI governance, explicitly addressing the use of artificial intelligence in elections. Compared to international experiences, which are often too broad or limited to either prohibiting or restricting deepfakes, the Brazilian model is notable for its normative breadth, institutional coordination, and attempt to hold platforms accountable. It combines repressive tools – such as content takedown mechanisms and political sanctions – with proactive obligations imposed on application providers to improve transparency, reduce harmful content, and promote electoral integrity.

Nonetheless, the Resolution also reveals structural challenges. In the absence of a general AI law in Brazil, it operates within a fragmented legal ecosystem, lacking a unified set of principles to ensure consistency across sectors. This regulatory gap means that important principles – such as algorithmic transparency, auditability, and human supervision – and other general norms regarding, for example, the use of sensitive personal data for content personalization and microtargeting, are not properly addressed by the Brazilian jurisdiction. Moreover, the Resolution's effectiveness depends on the judiciary's capacity to interpret and apply its provisions consistently. Without a robust and coherent jurisprudential framework, uncertainties will persist regarding the thresholds for content removal, the proportionality of sanctions, and the scope of platform duties.

Empirical evidence from the 2024 municipal elections – marked by several examples of the circulation of AI-generated disinformation targeting candidates – underscores the urgency of building institutional capacity, fostering cross-sectoral cooperation, and advancing public debate on the role of technology in democracy. The Resolution's success will ultimately depend not only on the legal norms it articulates but also on its integration into a broader ecosystem of digital governance, judicial engagement, and civil society oversight.

By offering a concrete, though still imperfect, response to the intersection of AI

and electoral law, Brazil's experience can serve as both a reference and a warning: while sectorial regulation may provide timely safeguards in high-risk domains, it cannot replace the need for comprehensive, systemic approaches to AI governance rooted in democratic values and constitutional principles.

REFERENCES

ARBEL, Yonathan A.; TOKSON, Matthew J.; LIN, Albert. Systemic Regulation of Artificial Intelligence (December 16, 2023). *AI Safety Legal Paper Series* 12-24, 56 Ariz. St. L. J. 545, University of Utah College of Law Research Paper No. 582, Available at SSRN: https://ssrn.com/abstract=4666854

BALDWIN, Robert; CAVE, Martin; LODGE, Martin. *Understanding Regulation*: Theory, Strategy, and Practice. 2nd ed. Oxford: Oxford University Press, 2012, chapter 2.

BOMMASANI, R. et al. *On the opportunities and risks of foundation models*. 2021. ArXiv:2108.07258.

Brazilian Superior Electoral Court (TSE). *Countering Disinformation Program for the 2022 Elections*. Available at: https://international.tse.jus.br/en/misinformation-and-fake-news/tse-brazil-counter-disinformation-program-2022-f.pdf.

Brazilian Superior Electoral Court (TSE). Eleições 2022: TSE e plataformas digitais firmam parceria para combate à desinformação. TSE, 07.02.2022. Available at: https://www.tse.jus.br/comunicacao/noticias/2022/Fevereiro/eleicoes-2022-tse-e-plataformas-digitais-firmam-parceria-para-combate-a-desinformacao

Brazilian Superior Electoral Court (TSE). *Resolution No. 23.714 of October 20, 2022*. Available at: https://www.tse.jus.br/legislacao/compilada/res/2022/resolucao-no-23-714-de-20-de-outubro-de-2022.

CAJUEIRO, Daniel O.; CELESTINO, Victor R. R. A *Comprehensive Review of Artificial Intelligence Regulation:* Weighing Ethical Principles and Innovation. 09.09.2024. Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4950727.

CARROZZINI, Alina. Shooting Democracy in the Foot? The Romanian Constitutional Court's Annulment of Presidential Elections. Verfassungsblog, 13 December 2024. Available at: https://verfassungsblog.de/shooting-democracy-in-the-foot/

CARVALHO, Lucas Borges de. A democracia frustrada: fake news, política e liberdade de expressão nas redes sociais. *Internet & Sociedade*, V.1, N.1 – 2020

CELESTE, Edoardo, *Digital constitutionalism: the role of internet bills of rights*, Abingdon, Oxon; New York, NY: Routledge, 2021.

China Law Translate. Provisions on the Administration of Deep Synthesis Internet Information Services. China Law Translate, 11.12.2022. Available at: <https://www.chinalawtrans-

late.com/en/deep-synthesis/>.

COHEN, Julie E. *Between truth and power*: The legal constructions of information capitalism. New York: Oxford University Press, 2019

COSSER, Kirsten. Will the real Slim Shady please stand up? Eminem video endorsing South African opposition party EFF and bashing the ruling ANC is a deepfake. Africa Check, 05.03.2024. Available at: <https://africacheck.org/fact-checks/meta-programme-fact-checks/will-real-slim-shady-please-stand-eminem-video-endorsing>.

DEL VALLE, Gaby. Political Consultant Behind the Joe Biden Deepfake Robocalls Faces $6 Million Fine. *The Verge*, May 23, 2024. Available at: https://www.theverge.com/2024/5/23/24163411/fcc-fine-biden-deepfake-robocalls-steve-kramer-lingo-telecom.

DEVINE, Curt; O'SULLIVAN, Donie; LYNGAAS, Sean. A fake recording of a candidate saying he'd rigged the election went viral. Experts say it's only the beginning. *CNN*, 01.02.2024. Available at: <https://edition.cnn.com/2024/02/01/politics/election-deepfake-threats-invs/index.html>.

ELLIOTT, Vittoria. The WIRED AI Elections Project. *Wired*, 30.05.2024. Available at: <https://www.wired.com/story/generative-ai-global-elections/>.

ELLUL, Joshua. Should we regulate AI or some uses of software. *Discov Artif Intell* **2**, 5 (2022). https://doi.org/10.1007/s44163-022-00021-9

European Partnership for Democracy (EPD). *Artificial Intelligence and Elections: Protecting Electoral Integrity in the Age of AI*. September 2024. Available at: https://epd.eu/content/uploads/2024/09/AI-and-elections.pdf.

FINOCCHIARO, Giusella. The regulation of artificial intelligence. *AI & Soc 39*, 1961–1968 (2024). https://doi.org/10.1007/s00146-023-01650-z. ELLUL, Joshua. Should we regulate AI or some uses of software. *Discov Artif Intell* **2**, 5 (2022). https://doi.org/10.1007/s44163-022-00021-9

FLORIDI, Luciano. Artificial intelligence, deepfakes and a future of ectypes. *Philosophy and Technology*. 2018;31:317–321. doi: 10.1007/s13347-018-0325-3

GIKAY, Asress Adimi. Risks, innovation, and adaptability in the UK's incrementalism versus the European Union's comprehensive artificial intelligence regulation, *International Journal of Law and Information Technology*, Volume 32, 2024, eaae013, https://doi.org/10.1093/ijlit/eaae013.

GREENWALD, Glenn. *No place to hide*: Edward Snowden, the NSA, and the Surveillance State. New York, NY: Metropolitan Books, 2014

GREGORIO, Giovanni de. *Digital Constitutionalism in Europe*: reframing rights and powers in the algorithmic society. Cambridge: Cambridge University Press, 2022

GREGORIO, Giovanni De; STREMLAU, Nicole, Internet Shutdowns in Africa | Internet Shutdowns and the Limits of Law, *International Journal of Communication,* v. 14, n. 0, p. 20, 2020.

HAO, Karen. How Facebook got addicted to spreading misinformation. *MIT Technology Review*, 11.03.2021. Available at: <https://www.technologyreview.com/2021/03/11/1020600/facebook-

responsible-ai-misinformation/>.

HILLMAN, Jonathan E., *The digital Silk Road*: China's quest to wire the world and win the future, First Edition. New York: Harper Business, 2021.

International Institute for Democracy and Electoral Assistance. *Artificial Intelligence for Electoral Management: A Guide for Electoral Management Bodies.* International IDEA, 2024. https://www.idea.int/sites/default/files/2024-04/artificial-intelligence-for-electoral-management.pdf

International Panel on the Information Environment [I. Trauthig, P. N. Howard, S. Valenzuela (eds.)]. *The Role of Generative AI Use in 2024 Elections Worldwide*. Zurich, Switzerland: IPIE, 2025. Technical Paper, TP2025.2, doi: 10.61452/HZUE9853.

JUDGE, Brian; NITZBERG, Mark; RUSSELL, Stuart. When code isn't law: rethinking regulation for artificial intelligence, *Policy and Society*, Volume 44, Issue 1, January 2025

KHAN, Lina. Amazon's Antitrust Paradox. *Yale Law Journal 126*, n. 3, jan. 2017, 710-805

KRÖGER, Jacob Leon; ERRENST, Emilia; NAU, Niklas; OJANPERÄ, Sanna, What is Political Microtargeting (May 22, 2024). This chapter is an excerpt from: Kröger, J., Errenst, E., Nau, N., and Ojanperä, S. (2024). *Mitigating the Risks of Political Microtargeting* – Guidance for Policymakers, Civil Society, and Development Cooperation, pp. 7-9. (forthcoming), Available at SSRN: https://ssrn.com/abstract=4856178 or http://dx.doi.org/10.2139/ssrn.4856178

LARKIN, CJ. Regulating Election Deepfakes: A Comparison of State Laws. Tech Policy Press, 8 January 2025. Available at: <https://www.techpolicy.press/regulating-election-deepfakes-a-comparison-of-state-laws/>.

LEHDONVIRTA, Vili. *Cloud empires: how digital platforms are overtaking the state and how we can regain control.* Cambridge, Massachusetts; London, England: The MIT Press, 2022

LIMA, Cristiano. A whistleblower's power: Key takeaways from the Facebook Papers. *The Washington Post,* 26.10.2021. https://www.washingtonpost.com/technology/2021/10/25/what-are-the-facebook-papers/.

MACHADO, Caio; KIRA, Beatriz; HIRSCH, Gustavo; MARCHAL, Nahema; KOLLANYI, Bence; HOWARD, Philip N.; LEDERER, Thomas; BARASH, Vlad. News and Political Information Consumption in Brazil: Mapping the First Round of the 2018 Brazilian Presidential Election on Twitter. Data Memo 2018. Oxford, UK: *Project on Computational Propaganda*, 2018.

MARQUES, Ana Luiza Pinto Coelho. *Erosão constitucional na era digital*: uma análise a partir do processo eleitoral brasileiro de 2018. São Paulo: Editora Dialética, 2024.

MATHEW, Binny; DUTT, Ritam; GOYAL, Pawan; MUKHERJEE, Animesh. Spread of hate speech in online social media. ArXiv:1812.01693, 2019. Available at: https://arxiv.org/abs/1812.01693.

MELLO, Patrícia Campos. Facebook não prioriza Brasil contra fake news, diz Frances Haugen, do Facebook Papers. *Folha de São Paulo*, 03.07.2022. https://www1.folha.uol.com.br/po-

der/2022/07/facebook-nao-prioriza-brasil-contra-fake-news-diz-frances-haugen-do-facebook-papers.shtml

MENN, Joseph. Russians made video falsely accusing Harris of hit-and-run, Microsoft says. The Washington Post, 17.09.2024. Available at: https://www.washingtonpost.com/technology/2024/09/17/russia-fake-videos-election-interference-kamala-harris/

MEYER, Emilio Peluso Neder. *Constitutional Erosion in Brazil*. Oxford; New York: Hart Publishing, 2021.

MEYER, Emilio Peluso Neder. BUSTAMANTE, Thomas. Accountability in Brazil: Between Democracy and the Digital Age." *Verfassungsblog*, May 8, 2023. https://verfassungsblog.de/accountability-in-brazil/

MOROZOV, Evgeny. *Big Tech:* a ascensão dos dados e a morte da política. São Paulo: Ubu Editora, 2018

MSC. A Tech Accord to Combat Deceptive Use of AI in 2024 Elections. MSC. Available at: <https://securityconference.org/en/aielectionsaccord/>.

MYERS, Steven Lee; THOMPSON, Stuart A. A.I. Is Starting to Wear Down Democracy. *The New York Times*, June 26, 2025. Available at: <https://www.nytimes.com/2025/06/26/technology/ai-elections-democracy.html>.

NARAYANAN, Arvind; KAPOOR, Sayash. *AI Snake Oil:* What Artificial Intelligence Can Do, What It Can't, and How to Tell the Difference. Princeton, NJ: Princeton University Press, 2024

NICAS, Jack; HERRERA, Lucía Cholakian. Las campañas electorales de Argentina recurren a la IA. *The New York Times*, 15.11.2023. Available at: <https://www.nytimes.com/es/2023/11/15/espanol/elecciones-argentina-imagenes-inteligencia-artificial.html>.

NICOLAU, Jairo. *O Brasil dobrou à direita:* uma radiografia da eleição de Bolsonaro em 2018. Rio de Janeiro: Zahar, 2020

NORDEN, Lawrence. HARRIS, David Evan. New Tech Accord to Fight AI Threats to 2024 Lacks Accountability for Companies. Brennan Center for Justice, 29.02.2024. Available at: https://www.brennancenter.org/our-work/analysis-opinion/new-tech-accord-fight-ai-threats-2024-lacks-accountability-companies

OECD. *Recommendation of the Council on Artificial Intelligence*. OECD/LEGAL/0449, 2019. Available at: https://oecd.ai/en/assets/files/OECD-LEGAL-0449-en.pdf>.

Public Citizen. Tracker: State Legislation on Deepfakes in Elections. Public Citizen. Available at: <https://www.citizen.org/article/tracker-legislation-on-deepfakes-in-elections/>.

RABITSCH, Armin; CALABRESE, Sofia. The EU's Artificial Intelligence Act and its Impact on Electoral Processes: a Human Rights-Based Approach. European Partnership for Democracy, September 2024. Available at: < https://epd.eu/content/uploads/2024/09/AI-and-elections.pdf>.

RADU, Roxana. Expert Comment: Which urgent tech lessons must the EU take from Romania's

election? Oxford News, 22 may 2025. Available at: <https://www.ox.ac.uk/news/2025-05-22-expert-comment-which-urgent-tech-lessons-must-eu-take-romania-s-election>.

ROBINS-EARLY, Nick. How did Donald Trump end up posting Taylor Swift deepfakes? The Guardian, 26.08.2024. Available at: https://www.theguardian.com/technology/article/2024/aug/24/trump-taylor-swift-deepfakes-ai

RUEDIGER, Marco Aurélio. GRASSI, Amaro. *Redes sociais nas eleições 2018*: Policy Paper 1. Rio de Janeiro: FGV DAPP, 2018

RUEDIGER, Marco Aurelio. GRASSI, Amaro (coord.). O ecossistema digital nas eleições municipais de 2020 no Brasil: o buzz da desconfiança no sistema eleitoral no Facebook, Youtube e Twitter. Policy Paper. Rio de Janeiro: FGV DAPP, 2020

SANTOS, João Guilherme dos. FREITAS, Miguel. ALDÉ, Alessandra. SANTOS, Karina. CUNHA, Vanessa. WhatsApp, política mobile e desinformação: a hidra nas eleições presidenciais de 2018. *Comunicação & Sociedade*. V. 41, n. 2, 2019

SCHULDT, Lasse. Every Fake You Make: Blanket Deepfake Bans Are the Next Level in Asia's War on Fake News. Verfassungsblog, October 9, 2024. Available at: https://verfassungsblog.de/fake-news-content-moderation-singapoer-south-korea-election/.

SIMCHON, Almog; EDWARDS, Matthew; LEWANDOWSKY, Stephan. The persuasive effects of political microtargeting in the age of generative artificial intelligence, *PNAS Nexus*, Volume 3, Issue 2, February 2024, pgae035, https://doi.org/10.1093/pnasnexus/pgae035

SOON, Carol. QUEK, Samantha. Safeguarding elections from threats posed by artificial intelligence. IPS Working Papers No. 56, August, 2024. Available at: https://lkyspp.nus.edu.sg/docs/default-source/ips/ips-working-paper-no-56_safeguarding-elections-from-threats-posed-by-artificial-intelligence.pdf

TAN, Bryan; KONG, Hannah; GOH, Eng Han. Singapore bans deepfakes in elections. ReedSmith, 04 November 2024. Avilable at: <https://www.reedsmith.com/en/perspectives/2024/11/singapore-bans-deepfakes-in-elections>.

TOKSON, Matthew J.; LIN, Albert. Systemic Regulation of Artificial Intelligence (December 16, 2023). *AI Safety Legal Paper Series* 12-24, 56 Ariz. St. L. J. 545, University of Utah College of Law Research Paper No. 582, Available at SSRN: https://ssrn.com/abstract=4666854.

TUFEKCI, Zeynep. The road from Tahrir to Trump. *MIT Technology review*, vol. 121. n. 05, set/out 2018

UNESCO. *Elections in Digital Times: A Guide for Electoral Practitioners.* Paris: UNESCO, 2023. https://unesdoc.unesco.org/ark:/48223/pf0000382102

United Nations. Can artificial intelligence (AI) influence elections? United Nations, 07.06.2024. Available at: <https://unric.org/en/can-artificial-intelligence-ai-influence-elections/>

United Nations. UN Warns of Disinformation Threat to 2024 Electoral Cycle, Global Stability. *UN*

News, September 19, 2024. https://news.un.org/en/story/2024/09/1154316.

VAIDHYANATHAN, Siva. *Antisocial Media:* How Facebook Disconnects Us and Undermines Democracy. 1ª ed. Oxford University Press, 2018

VOSOUGHI, Soroush; ROY, Deb; ARAL, Sinan. The spread of true and false news online. Science, VOL. 359, NO. 6380

WENDLING, Mike. "Trump sides with tech bosses in Maga fight over immigrant visas", *BBC News*, 24 Dec. 2024, https://www.bbc.com/news/articles/clyv7gxp02yo.