Editors

LUCA BELLI and WALTER B. GASPAR

# THE QUEST FOR AI SOVEREIGNTY, TRANSPARENCY AND ACCOUNTABILITY

## Official Outcome of the UN IGF Data and Artificial Intelligence Governance Coalition

Preface by
ANA BRIAN NOUGRÈRES

Preliminary version presented at the United Nations Internet Governance Forum 2023

IGF Internet Governance Forum

FGV DIREITO RIO

# The Quest for AI Sovereignty, Transparency and Accountability

## Official Outcome of the UN IGF Data and Artificial Intelligence Governance Coalition

*Luca Belli* and *Walter B. Gaspar*
Editors

**This volume is the preliminary version of the outcome report presented at the United Nations Internet Governance Forum (IGF), in Kyoto, Japan, in October 2023, to receive feedback from the IGF community.**

This volume is the result of a participatory process developed by the Data and Artificial Intelligence Governance (DAIG) Coalition of the United Nations Internet Governance Forum (IGF). The views and opinions expressed in this volume are those of the authors and do not necessarily reflect those of the United Nations Secretariat. The designations and terminology employed may not conform to United Nations practice and do not imply the expression of any opinion whatsoever on the part of the Organization. For any comments on the chapters of this volume, please contact the authors or the editors.

# The Quest for AI Sovereignty, Transparency and Accountability

## Official Outcome of the UN IGF Data and Artificial Intelligence Governance Coalition

*Luca Belli* and *Walter B. Gaspar*
Editors

**This volume is the preliminary version of the outcome report presented at the United Nations Internet Governance Forum (IGF), in Kyoto, Japan, in October 2023, to receive feedback from the IGF community.**

This volume is the result of a participatory process developed by the Data and Artificial Intelligence Governance (DAIG) Coalition of the United Nations Internet Governance Forum (IGF). The views and opinions expressed in this volume are those of the authors and do not necessarily reflect those of the United Nations Secretariat. The designations and terminology employed may not conform to United Nations practice and do not imply the expression of any opinion whatsoever on the part of the Organization. For any comments on the chapters of this volume, please contact the authors or the editors.

This material, its results and conclusions are the responsibility of the authors
and do not represent, in any way, the institutional position of the
Getulio Vargas Foundation / FGV Direito Rio.

# CONTENTS

## PART 4

## PART 5

# PREFACE

## The Promise of a New Coalition Working on Artificial Intelligence, Transparency and Accountability

**Ana Brian Nougrères, UN Special Rapporteur on the right to privacy**

The Coalition on Data and Artificial Intelligence Governance (DAIG) is a multistakeholder group that was established under the auspices of the United Nations Internet Governance Forum (UN IGF). The idea of the establishment of such a unique group emerged as an outcome of the Data Governance School LatAm 2023, organized by the Center for Technology and Society at FGV Law School, Rio de Janeiro.

The DAIG Coalition aims at fostering analyses of existing approaches to data and artificial intelligence governance, promoting the debate of shared problems and potential aimed at achieving sustainable and effective models of data and artificial intelligence governance. For that purpose, the group promotes collective studies and multistakeholder interactions to collect and discuss evidence, critically analyze existing regulatory and institutional arrangements, and propose policy updates in artificial intelligence.

The DAIG was created to act as a hub to connect global UN IGF discussions with regional and local initiatives, with a specific focus on Global South debates. Hence, it invited researchers and practitioners to submit papers about artificial intelligence transparency and accountability. These papers formed the first **Annual Report of the UN DAIG Coalition** that has the intention of providing reflections on artificial intelligence transparency and accountability, and exploring the new concept of artificial intelligence sovereignty.

This work brings much needed thematic analyses from several countries and regions (Argentina, Brazil, China, Colombia, India, Mexico, Nigeria, South Africa, the Global South, Europe, etc.) offering diverse perspectives on critical AI issues.

This Outcome Report teaches us the importance of initiatives such as the Artificial Intelligence Act (AIA) as a milestone in artificial intelligence governance and, at the same time, raises concerns about the limited application of accountability requirements and the potential for vested interests to influence evaluations. This report shows us the intricate relationship between trade secrecy, intellectual property, and data subject's rights, how transparency must operate between them and how their special relationship challenges the authorities. It exposes the importance of cooperation and how international efforts can effectively promote transparency and accountability.

Importantly, this report analyzes how a globally harmonized, transparent, accountable artificial intelligence landscape is necessary, balancing innovation and responsible governance. It explains how to mitigate risks with the implementation of targeted transparency and accountability. It investigates the challenges of opaque artificial intelligence systems, the importance of transparency and accountability for the responsible development and deployment of artificial intelligence systems and how technical tools can help to create systems that are comprehensible, ethical, and accountable.

The concept of responsible artificial intelligence appears in several papers. Other papers encourage, as well, ethical practices in decision-making as a contribution to a sustainable framework for artificial intelligence. The triple dimension of algorithmic transparency (traceability, explainability and auditability) is also highlighted, as well as the importance of citizen participation and collaboration in the process of the design of artificial intelligence systems.

Country experiences referring to sandboxes in the public sector, emphasizing principles in national strategies, analyzing regulatory frameworks, mapping harms and impacts are developed on several papers and contribute with practical examples to the more abstract view provided by other authors.

Special attention is given to the role of civil society organizations, which ensure fairness, transparency, and accountability, and also facilitate ethical governance of artificial intelligence for the good of the public.

A paper proposing artificial intelligence by corporate design refers to a way of preventing a risky adoption of artificial intelligence into corporate structures and presents a pragmatic solution for a responsible integration, upholding at the same time respect for the relevant ethical, legal, and algorithmic instances.

The impact of artificial intelligence and neurotechnologies, specially in what refers to immersive technologies and vulnerable sectors of the population shows the need to establish regulatory criteria to clarify international standards on new advances in science and technology, with a human rights point of view.

The conceptual framework for AI supply chain regulation focuses on principles of transparency, incentivization, efficacy, and accountability. It requires the use of various transparency mechanisms to enable critical information flow and modes of redress up and down an AI systems supply chain. The advent of general-purpose AI systems like OpenAI's GPT-4 complicates the challenge of allocating responsibility.

Factors such as who is designing them, how they are released, and what information is made available about them may impact the allocation of responsibilities for addressing potential risks are not easy to solve. The paper proposes that policymakers should focus on how artificial intelligence systems are released into public use to inform the allocation of responsibilities for addressing harms throughout an identified supply chain.

The risks and advantages associated with artificial intelligence applications in military operations, focusing on the role of artificial intelligence in enhancing Counter Unmanned Aircraft Systems (C-UAS) is also studied. The goal of this discussion is to inform on the potential added value, limitations, and ways to mitigate the risks of deploying artificial intelligence applications in military operations. This paper concludes that if the impact on democratic societies is not clarified and new requirements are not used on a tactical and conceptual level, artificial intelligence applications may not be deployed responsibly or lawfully.

The contributions featured in this volume invite further steps that can be taken by developing verification and validation of new requirements in real-life environments, and presenting these findings to decision-makers and stakeholders for reshaping legislation, certification, and policy guidelines.

As said, this useful Outcome Report is the product of the reflections of members of the DAIG Coalition, its chapters are dedicated to some of the most important issues that are of our concern in what refers to artificial intelligence, transparency and accountability. This work constitutes an important invitation to keep on working proactively towards a world in which the achievements of science and technology contribute as a catalyst for human rights and freedoms, always striving to focus on a human centric perspective.

# ABOUT THE AUTHORS

**Luca Belli**, PhD is Professor of Digital Governance and Regulation at Fundação Getulio Vargas (FGV) Law School, where he heads the Center for Technology and Society (CTS-FGV) and the CyberBRICS project, and associated researcher at Centre de Droit Public Comparé of Paris 2 University. He is co-founder and co-coordinator of the IGF Data and Artificial Intelligence Coalition (DAIG), Editor of the International Data Privacy Law Journal, and Board Member of the Global Digital Inclusion Partnership. Before joining FGV, Luca worked as an agent for the Council of Europe Internet Governance Unit and served as a Network Neutrality Expert for the Council of Europe. He is author of more than 50 academic publications which have been quoted by numerous media outlets, including The Economist, Financial Times, Forbes, Le Monde, BBC, The Hill, China Today, O Globo, Folha de São Paulo, El Pais, and La Stampa. Luca holds a PhD in Public Law from Université Paris Panthéon-Assas and can be found on LinkedIn and on Twitter as @1lucabelli.

**Ana Brian Nougrères**, PhD is the United Nations Special Rapporteur on the right to privacy and she took up the mandate on 1 August 2021. A Professor of Law, Privacy and ICT at the School of Engineering, University of Montevideo and a Professor of Law, Data Protection and ICT at the School of Law, University of the Republic, Montevideo. She is based in Uruguay and is a practicing Attorney-at-law and Consultant on data protection. She presented her first UN report Privacy and personal data protection in Ibero-America: a step towards globalization? A/HRC/49/55 at the Human Rights Council in March 2022.

**Walter Britto Gaspar**, lawyer, graduated Law at FGV in 2015. Master in Public Health at UERJ (2017), studying the interface between innovation, intellectual property and access to medicines policies in Brazil. Grantee of the Fundación Botín Programme for the Public Interest in Latin America (2013). Researcher in the Fiocruz and Shuttleworth Foundation project on intellectual property and access to medicines (2017). National Coordinator of the NGO Universities Allied for Essential Medicines (2013-2016). Certified Graphic Designer by the Istituto Europeo di Design (2018). Author of “O que é Creative

Commons? Novos modelos de direito autoral em um mundo mais criativo". Currently, researcher in the CyberBRICS project at FGV's Center for Technology and Society and Ph.D. student at the Public Policies, Strategies and Development Programme at the Economics Institute of at he Federal University of Rio de Janeiro (UFRJ).

**Bhoomika Agarwal** is a Research Associate with The Dialogue. She completed her BA LLB (H) from Guru Gobind Singh Indraprastha University. Her focus areas include Tech Policy and Competition Law.

**Bhavya Birla** is a Research Associate at The Dialogue. During his tenure, he has worked in the fields of AI, Digital Economy, Data Privacy, and Telecommunications policy, informing government policies and advocating for academically backed policies. He is an advocate for individual privacy and responsible development of AI.

**Pedro Braga**, DSc student in Systems and Computer Engineering programme at the Federal University of Rio de Janeiro (UFRJ) and researcher of the department of Law and GovTech at the Institute for Technology & Society (ITS Rio). He is interested in science-technology-society studies, free software and participatory software development. He has a Master's degree in History of Sciences and Techniques and Epistemology (HCTE/UFRJ) and a Bachelor's degree in Computer Science (UFRJ). He is also a lecturer on the Systems Analysis and Development undergraduate course at Faculdade Anhanguera, São João de Meriti campus (RJ).

**Ian Brown**, PhD is a consultant on Internet regulation, particularly relating to information security and privacy, digital elements of the election cycle, and pro-competition mechanisms such as interoperability. He is a visiting professor at the Centre for Technology and Society at Fundação Getulio Vargas (FGV) Law School in Rio de Janeiro, and an ACM Distinguished Scientist. He was previously Principal Scientific Officer at the UK government's Department for Digital, Culture, Media and Sport; Professor of Information Security and Privacy at the University of Oxford's Internet Institute; and a Knowledge Exchange Fellow with the Commonwealth Secretariat and UK National Crime Agency.

**Jesús Javier Sánchez García** holds a degree in International Relations from the National Autonomous University of Mexico, completed a Diploma in Privacy, Regulation and Data Governance from the Center for Economic Research and Teaching, A.C. and a Diploma in International Cooperation from the Dr. José Luis María Mora Research Institute. He has more than 10 years of professional experience in the field of International Relations with close links to issues such as protection of personal data, privacy and access to information, currently he is working at the National Institute of Transparency, Access to Information and Personal Data Protection (INAI Mexico).

**María Julia Giorgelli** is an independent researcher. The document reviews the regulatory framework on artificial intelligence (hereinafter AI) in Argentina. It also provides background information on the international commitments endorsed by the country, lists various actions taken by the National Executive Branch and summarizes the latest bills submitted at the national level. In all cases, the focus is on the right to privacy/personal data and transparency/information.

**Juan David Gutiérrez** is Associate Professor at the Alberto Lleras Camargo School of Government, Universidad de los Andes. PhD in Public Policy from the School of Government at Oxford University. His research interests include public policy, artificial intelligence, and natural resource governance.

**Divij Joshi** is a lawyer and researcher studying the intersections of technology, regulation and society, based in London and India. He is a Doctoral Researcher at University College London. His research examines legal and political implications of platforms and information Infrastructures. He is a Visiting Fellow at the NCBS Archives, a former Mozilla Tech Policy Fellow and edits the SpicyIP Blog.

**Jonathan Mendoza Iserte** holds a Ph.D. and a master's degree in Law from the National Autonomous University of Mexico. He also has a master's degree in General Data Protection Regulation of the European Union from the National Distance Education University of Spain and a certificate for the specialization course “Cybersecurity Summer Bootcamp – Policy Makers,” organized by he University of León, Spain, and the National Institute of Cybersecurity (INCIBE). Currently, Dr. Mendoza is the Secretary of Personal Data Protection at the National Institute of Transparency, Access to Information, and Personal Data Protection (INAI Mexico).

**Natalia L. Monti** is a Lawyer, Master in Human Rights, member of the Kamanau Foundation, drafter of the project before the Inter-American Juridical Committee of the OAS on the Inter-American Principles on Neurosciences, Neurotechnologies and Human Rights. She is also a member of the Center for the Protection of Personal Data of the Ombudsman's Office CABA, Argentina.

**Thiago Moraes** is a PhD candidate of Vrije Universiteit Brussels (VUB) and Universidade de Brasilia (UnB). He works as the Coordinator of Innovation and Research at the Brazilian Data Protection Authority (ANPD), and is also Co-founder and Counselor of the Laboratory of Public Policy and Internet (LAPIN).

**Sizwe Snail ka Mtuze**, admitted Attorney of the South African High Court Sizwe Lindelo Snail ka Mtuze holds an LLB from the University of Pretoria. He is a practising attorney with the law firm Snail Attorneys at Law. He holds an LLM from the University of South Africa. Sizwe is registered with the University of Fort Hare for an LLD. Sizwe is an adjunct Professor at the Nelson Mandela University, Mercantile Law Department (2020- to date) and has been a Research Fellow since 2014 as well as Lecturer in the field of cyberlaw at the University of Fort Hare since (2018- date) Sizwe is also a Visiting Professor at CTS-FGV University, Rio de Janeiro since 2022 . Sizwe was a member of the South African Information Regulator (IR) from (2016-2021). Sizwe served on the National Cyber Security Advisory Council of the DTPS. (2014 -2016) He was deputy chair and chair for the Law Society of South Africa (LSSA) E-Law Committee (2013-to 2021). He has been an Advisory Member of the Cyber BRICS, FGV Rio de Janeiro, Brazil (2019 to date ) Sizwe was the co-editor and author of CyberlawSA III: The Law of the Internet in South Africa (2012) but also CyberlawSA IV: The Law of the Internet in South Africa published in 2022.

**Sarah Muñoz-Cadena** is a student of the Master in Economics of Public Policy at Universidad del Rosario and researcher at Policéntrico. Political scientist and professional in Government and Public Affairs with complementary studies in journalism at Universidad de los Andes. She researches on governance of artificial intelligence and design thinking.

**Melody Musoni**, LLB (WITS), LLM (WITS), LLD (WITS) – Policy Officer at the European Centre for Development Policy Management, The Netherlands. Before joining ECDPM, Melody worked as a data protection senior expert and advisor at the Southern African Development Community (SADC) Secretariat. Melody has a decade long experience in both legal practice and academia where she specialised in privacy law, cybersecurity, and information technology law. Melody recently finished her PhD degree in cybercrime law and cloud computing law. She holds a master of laws degree and a bachelor of laws degree from the University of Witwatersrand.

**Nicola Palladino**, (PhD in Sociology, Social Analysis and Public Policy) is Research Fellow at the Trinity College Dublin’s Long Room Hub Arts and Humanities Research Institute, where is carrying out the research project “Hybrid Governance for Trustworthy and Human-Centric Artificial Intelligence: From Principles to Practices” funded from the European Union’s Horizon 2020 Research and Innovation Programme under the HUMAN+ COFUND Marie Skłodowska-Curie. He studied and worked at the University of Salerno and served as postdoctoral researcher at the School of Law and Government of the Dublin City University. He is also a member of the Digital Constitutionalism Network supported by the Center of Advanced Internet Studies in Bochum, Germany. His main research interests include Global Internet Governance, Digital Policies, AI Ethics and Regulation. He recently published the volumes “Legitimacy, Power, and Inequalities in the Multistakeholder Internet Governance: Analyzing IANA Transition” and “The Content Governance Dilemma: Digital Constitutionalism, Social Media and the Search for a Global Standard” for the Palgrave Information Technology and Global Governance book series.

**Smriti Parsheera** is a PhD candidate at the Indian Institute of Technology Delhi. Her research interests include privacy, data governance and digital rights. Until recently, she was the India Fellow at the CyberBRICS Project. Before that she led the technology policy vertical at New Delhi’s National Institute of Public Finance and Policy. She has also worked for the Competition Commission of India and UNDP India. Smriti studied law at the National Law School of India University, Bangalore and obtained her LLM from the University of Pennsylvania with a Certificate in Law and Business from the Wharton School. She recently edited the book Private and Controversial: When Privacy and Public Health Meet in India that was published by HarperCollins.

**Christian Perrone** is a highly accomplished legal scholar with a Ph.D. in International Law and Digital Law from UERJ and an LL.M. in International Law from the University of Cambridge, UK. He also holds a Diploma in International Human Rights Law from the European University Institute (EUI, Italy) and was a Fulbright Scholar at Georgetown University, USA. Perrone has extensive experience in international law and human rights, having served as Secretary of the Inter-American Juridical Committee of the Organization (IAJC) of American States (OAS) and worked as an expert with the Inter-American Commission on Human Rights and the Inter-American Court of Human Rights. He is currently a Public Policy Consultant at ITS, where he heads the organization's Law and Technology and GovTech departments.

**Nadia Elsa Gervacio Rivera** holds a Master degree in Education and a bachelor degree in Translation and Interpretation with a specialization in Technical and Scientific translation by the American Technological University in Mexico City. She has more than 10 years of professional experience in the public and private sectors and currently she is working at the National Institute of Transparency, Access to Information and Personal Data Protection (INAI Mexico).

**Rama Vedashree** is a renowned tech-policy expert with over 35 years of experience in the Technology Industry with stints in NIIT Technologies, Microsoft India and NASSCOM. She retired as CEO of Data Security Council of India (DSCI).

**Kamesh Shekhar** is a Programme Manager, leading the verticals on Data Governance and privacy at The Dialogue. He was also a Fellow at The Internet Society. His area of research covers informational privacy, surveillance technology, intermediary liability, safe harbour, issues of mis/disinformation on social media, AI governance etc.

**Jameela Sahiba** is a Senior Programme Manager at The Dialogue leading the vertical on Emerging Tech in addition to building and managing parliamentary engagement and outreach. She is a lawyer by profession and her interest areas include understanding how emerging technologies like AI/ML will influence various public sectors and how regulatory frameworks will take shape to govern that. Prior to this, she worked as Chief of Staff for the office of Dr. Amar Patnaik, Member of Parliament (Upper House), India.

**Attamongkol Tantratian** is a Doctor of Juridical Science (S.J.D.) candidate at Indiana University Maurer School of Law, USA, where he serves as a graduate affiliate for the IU Center of Applied Cybersecurity Research. Additionally, he is a tech policy fellow at a Boston-based think tank Alliance for Citizen Engagement, where he researches and publishes non-partisan policy briefs on pressing technology regulation issues. Before commencing his doctoral studies, Atta was a policy researcher at Thailand Development Research Institute, where he studied Thailand's first data protection legislation and drafted compliance guidelines for different industries. Atta earned an LL.M. from Indiana University and his LL.B. from Thammasat University in Thailand. He also holds privacy certifications CIPP/E and CIPP/US from IAPP.

**Rolf H. Weber**, PhD is professor of international business law at Zurich University acting there as co-director of the Center for Information Technology, Society, and Law (ITSL) and as co-director of the Blockchain Center. In parallel, Prof. Weber is a practicing attorney-at-law in Zurich. In 2003/05, he was involved in the IGF implementation and since then regularly participates in the IGF. Frequently, Prof. Weber publishes and speaks on Internet- and Blockchain-related legal issues as well as on topics of international finance and business law. His main fields of research and practice are IT- and Internet, Blockchain, finance, international business and competition law. He is fluent in German, English and French.

**Wayne Wei Wang** is now a Ph.D. Candidate in Law and Technology at the University of Hong Kong (HKU), and a Fellow-in-Rio at FGV Rio Law School (FGV Direito Rio) in Brazil. Trained in Engineering and Law, Wayne focuses his research interests on Intellectual Property, Data Protection, Algorithmic Governance, and S&T Studies, with a special focus on Law, Innovation, Technology, and Entrepreneurship in the Automating Global South. He (has) also held academic affiliations with universities and institutions in the USA, the UK, Germany, Poland, and Singapore. Prior to his Ph.D. studies, Wayne completed his LLM in Intellectual Property with Dean's Scholarship jointly conferred by the World Intellectual Property Organization (WIPO) and Queensland University of Technology (QUT) in Australia. He also worked as Data Analyst at a legal technology start-up in Shenzhen. Wayne graduated with his Double Bachelors in Engineering and Law as well as MPhil from Huazhong University of Science and Technology (China), with China National Scholarship from the Ministry of Education of P.R.China and University Outstanding Graduate Award.

**Yue Zhu** is an Assistant Professor at the Tongji University School of Law. His research focuses on privacy and data protection, as well as the regulation of emerging technologies. He previously worked for a leading international platform in China, where he gained practical experience in data privacy and AI compliance. Mr. Zhu's recent research covers the legal-technical analysis of privacy-enhancing technologies, digital watermarking and data protection, accessible protection of digital rights, the history of AI ethics, and EU digital legislation. He holds a bachelor's degree in economics from Peking University and a Juris Doctor from Washington University in St. Louis. Mr. Zhu has been involved in the drafting team of the (Chinese) Model Artificial Intelligence Law (Expert Draft Proposal) since 2023, whose English version 1.1 was launched at the Chinese Academy of Social Sciences.

# PART 1

## Framing the AI Sovereignty Debate

# 1 AI Transparency, AI Accountability, and AI Sovereignty: An Overview

**Luca Belli, Professor and Coordinator, Center for Technology & Society at FGV Law School.**

**Walter B. Gaspar, Researcher, Centre for Technology and Society at FGV Law School.**

## Abstract

This chapter presents the fundamental assumption of this volume, which is to acknowledge the transformative impact that AI systems can have and engage with the various existing frameworks that can regulate such systems. Particularly, this paper discusses the structure of this book, arguing that discussions on transparency, accountability, and sovereignty regorging AI governance, play a key role shaping the development of AI systems towards paths that align with the protection of fundamental rights, the respect of existing legal requirements, the promotion of sustainability and, ultimately, the maximisation of the public interest. In doing so, the chapter provides an overview of how the works collected in this volume trace the discussion of AI transparency, accountability, and sovereignty. The chapter acknowledges that these concepts have been given much attention throughout policy debates and have been defined in heterogeneous fashion, depending on the different public discourse arenas, and stakeholder interpretations.

## 1.1 The IGF Coalition on Data and Artificial Intelligence Governance

This volume marks the beginning of activities of the Dynamic Coalition on Data and Artificial Intelligence Governance (DAIG), a multistakeholder group established under the auspices of the United Nations Internet Governance Forum (IGF). The Coalition aims at fostering discussion of existing approaches to data and AI governance, promoting analysis of good and bad practices to identify what solutions should be replicated and which ones should be avoided by stakeholders to achieve a sustainable and effective data and AI governance.

To do so, the DAIG Coalition aims at promoting collective studies and multistakeholder interactions to collect and discuss evidence, critically analyse existing regulatory and institutional arrangements, and propose policy updates in the areas of AI and data governance. Importantly, the DAIG aims at acting as a hub to connect global UN IGF discussions with regional and local initiatives, with a particular focus on Global South debates.

This first Annual Report of the Coalition aims at fostering reflections on transparency, accountability, and sovereignty in the context of AI governance, with a particular focus on experiences of Global South countries, to provide valuable contributions that could feed into IGF 2023 discussions. Particularly, this volume aims at answering pressing questions on the governance and regulation of AI systems, which are likely to have an enormous impact on the evolutions of our societies, economies, and democracies.

The fundamental assumption of this volume is to acknowledge the transformative impact that AI systems can have and engage with the various existing frameworks aimed at directing the development of such systems. In this perspective, the discussion of the vaguely defined concepts of transparency, accountability and sovereignty is essential to provide further clarity to the AI governance debate and align it with the full protection of fundamental rights, the respect of existing legal requirements, the promotion of sustainability and, ultimately, the maximisation of the public interest.

To do so, this book seeks to present a diverse set of views, in the spirit of multistakeholder debate, from various sectors, countries, disciplines and theoretical backgrounds. Conspicuously, the works collected in this volume trace a picture of current discussions regarding transparency, accountability, and sovereignty, concepts that have been given much attention – and varying definitions in different contexts – throughout policy debates and, increasingly, within many public discourse arenas. The contributions presented in this volume are organised around five thematic axes, briefly exposed in the following sections.

## 1.2 Framing the AI sovereignty debate

The first part of this volume is dedicated to "Framing the AI sovereignty debate", a general theoretical framework for the connection between issues concerning the development, implementation and regulation of the various components of AI (eco)systems and matters of autonomy and self-determination are expanded upon. Luca Belli's "Exploring the Key AI Sovereignty Enablers (KASE) of Brazil, to Build an AI Sovereignty Stack" opens the debate indicating what is to be a running characteristic of many contributions: their focus on actionable, pragmatic frameworks.

Belli's proposed analytical tool is rooted in the concept of AI Sovereignty – "the capacity of a given country to understand, muster and develop AI systems, while retaining control, agency and, ultimately, self-determination over such systems". This frames the discussion under a complex web of geopolitical, sociotechnical, and legal considerations, whose core elements compose the AI Sovereignty Stack. The proposed perspective is attentive to the power dynamics involved in being a developer or an importer of transformational technologies such as AI, and venture into the governance and regulatory framework that can enable AI Sovereignty.

The first chapter is followed by responses where the KASE framework is put to the test, applying it to the South African and Indian environment. In "An Assessment of the Key AI Sovereignty Enablers within the South African context", Melody Musoni and Sizwe Snail argue that African countries are taking steps in carving out their position as competitors in the development of Artificial Intelligence (AI). The paper assesses the Key AI Sovereignty Enablers (KASE) framework proposed by Belli within the South African context. The paper provides recommendations on the way forward reading KASE in South Africa.

Subsequently, Divij Joshi's "AI Sovereignty in India – A Response to the KASE Framework" examines Indian AI policy and governance from the lens of Belli's 'Key Enablers of AI Sovereignty'. Further, the paper interrogates the potential and limitations of sovereignty-based discourses and frameworks, and examines how it might include questions of injustice, equity and democratic participation.

Importantly, both contributions discuss the interest of the framework as a retrospective and prospective tool of policy analysis, while highlighting the need for further work toward addressing the structural dimensions of the AI stack concerning market concentration, labour practices and community needs.

## 1.3 What Do AI Transparency and AI Accountability Mean?

The second part of this book is dedicated to "What Do AI Transparency and AI Accountability Mean?" and is opened by Rolf Weber's theoretical analysis of regulatory models. The chapter provides a deep dive into the concepts of transparency and accountability and how they fall short as countermeasures to AI's negative impacts. The author argues that such concepts should be complemented by auditability and observability.

The following chapters in this part concern various theoretical approaches toward providing responses to AI transparency and accountability concerns. Ian Brown's "A conceptual framework for AI supply chain regulation" discusses how policymakers and regulators can apply different responsibilities in the regulation of AI systems to their constituent parts. This approach complements requirements from a range of existing legal frameworks including data protection, copyright, equality and non-discrimination, and contractual liability.

The following chapter analyses how principled approaches can be translated into actionable measures. In "GenAI and the Goblet of Compliance: Delving into the Pensieve of Privacy Principles", Pranav Tiwari and colleagues discuss how a comprehensive privacy compliance framework for Generative AI can be created through multistakeholder cooperation, proposing sixteen key privacy principles tailored for Generative AI platforms.

The last chapter of this part is "Towards Trustworthy AI: Guidelines for Operationalisation and Responsible Adoption" by Jameela Sahiba and colleagues. This paper serves the purpose of converting the widely accepted principles of trustworthy AI into tangible, actionable steps designed for both AI developers and AI users, while offering a comprehensive approach that addresses both the technical and non-technical dimensions.

## 1.4 Western Perspectives on AI Governance

Part 3 explores "Western Perspectives on AI Governance", detailing various aspects of current debates in AI regulation in existing frameworks. In "The blind watcher: accountability mechanism in the Artificial Intelligence Act", Nicola Palladino conceptualizes accountability in the European AI Act and goes into detail on its risk-based approach to building trust in AI-powered settings, providing a critical view to how real-world institutional architectures can – or cannot – achieve the intended purpose.

Atta Tantratian presents a matter at the crux of the implementation of AI systems and personal data processing – the transparency of AI-generated inferences. In situations where data subject rights and trade secret law are at an impasse, the author considers that authorities should carefully strive to eliminate abuses of trade secret law that might harm transparency and the realization of data subject rights.

Giuseppe Cicu's "AI by Corporate Design" framework attempts to fit together business management and AI governance concerns into a step-by-step implementation process, from strategic planning to optimization. Cicu's effort brings AI transparency and accountability matters into the organizational structure of the firm, providing a game plan for responsible AI implementation.

The final chapter of this part concerns a particular field of AI applications with high potential impact and dire need for academic debate. In "Clarifying Military Advantages and Risks of AI Applications via a Scenario", Liisa Janssens provides a focused responsible AI framework for military applications, developed through a scenario-setting methodology for considering AI regulation's virtues and shortcomings. The disruptive nature of AI is considered in face of the demands of Rule of Law mechanisms to trace the requirements that make up responsible use of AI in military theaters.

## 1.5 Asian and African Perspectives on AI Governance

Part 4, "Asian and African Perspectives on AI Governance" is the first of two parts dedicated to exploring the connection between theory and practice of AI, looking at ongoing efforts in Global South countries. This part is opened by Wei Wang and Yue Zhu's chapter on "Operationalizable Accountability of (Generative) AI: Towards the Chinese AI Law?". The paper elucidates the disparate conceptualizations of AI accountability among various stakeholders at the Chinese level, thereby facilitating an informed discussion about the ambiguity and implementability of normative frameworks governing AI, specifically, regarding Generative AI.

Subsequently, Smriti Parsheera's "Seeking Policy, Technical and Operational Transparency in AI Systems: A Case Study of India's Digi Yatra Project" discusses the why and how of transparency obligations, as articulated in the AI governance discussions in India and in select international principles. It argues that the need for transparency permeates through the lifecycle of an AI project and identifies the policy layer, the technical layer and the operational layer as the key sites for fostering the transparency in AI projects.

In their paper on "Principles for Enabling Responsible AI Innovations in India: An Ecosystem Approach", Shekhar et al. argue for a principle-based approach coupled with a detailed classification of AI harms and impacts. The paper proposes a detailed multistakeholder approach which resonates with the foundational values of responsible AI envisioned by various jurisdictions geared towards ensuring that AI innovations align with societal values and priorities.

Lastly, Michael Karanicolas' chapter on "Developing AI standards to Serve the Majority World" argues for the need to develop AI standards beyond the "auspices of a handful of powerful regulatory blocs", and calls for the inclusion of the Majority World into standard-setting processes in international fora.

## 1.6 Latin American Perspectives on AI Governance

Part 5, "Latin American Perspectives", are testaments to the effort to approach the subjects analysed in this volume from a holistic viewpoint and an inclusive perspective. Thiago Moraes' paper on "Regulatory Sandboxes as Tools for Ethical and Responsible Innovation of Artificial Intelligence and their Synergies with Responsive Regulation" explores the role of regulatory sandboxes as tools to foster ethical and responsible innovation in AI systems and discusses the synergies of sandboxes with responsive regulatory theory. The analysis is carried out through bibliographical research with focus on experiences from the Global South (Brazil, Colombia and Singapore) and European countries.

Braga and Perrone's "AI Development Model for the Brazilian Justice Ecosystem" describes the use of machine learning techniques to amplify the analysis of judicial data and propose mechanisms to develop an AI for public policy. The paper highlights the successful implementation of an Operational AI Sandbox approach, ensuring the responsible development of technology in the public sector.

Subsequently, Gutiérrez and Muñoz-Cadena's "Building a repository of public algorithms: Case study of the dataset on automated decision-making systems in the Colombian public sector" documents how the team of scholars built the new repository of public algorithms in Colombia and describes how the data was collected, processed, and organized. The article also explains the main difficulties that the researchers encountered as well as the solutions that were implemented.

Lastly, the three Spanish-language chapters in the volume provide descriptive accounts of ongoing efforts to promote transparency and accountability. First, García, Rivera & Mendoza Iserte explore current and prospective frameworks in the international sphere, providing a detailed account and proposal of a path forward for Latin America. María Julia Giorgelli follows with a comprehensive account of the complex normative outlook for AI governance in Argentina, a country with a long trajectory in data governance and personal data protection regulation, and connects it with regional development.

Natalia Monti's chapter explores the cutting-edge issue of regulating AI in the context of neurotechnologies, and the unprecedented ethical and legal issues that arise. Particularly, Monti discusses the process leading to the Organization of American States' recent publication of the "Interamerican principles neurosciences, neurotechnologies and human rights" alongside a case study of the Chilean Supreme Court, which issued a landmark ruling on neurotechnology devices.

## 1.7 Conclusions

To conclude this chapter, we would like to stress that most of the contributions featured in this volume have an exploratory nature, venturing into a field that is acquiring enormous relevance, with the purpose of applying established or emerging conceptions of transparency, accountability, and sovereignty. To have meaningful application, these concepts need clear specification and experimentation.

While the research body dedicated to AI governance and regulation is in continuous expansion, many of the core elements utilised in the proposed frameworks are frequently vague. In this perspective, the aim of this volume is to offer a valid contribution to the study of how AI systems could be framed, stressing the fundamental goal of data governance, and calling stakeholders to engage into a much-needed collaborative effort, able to steer the evolution of AI in a sustainable fashion, guaranteeing that innovation and development meet the challenge of rule of law and democracy.

# 2 Exploring the Key AI Sovereignty Enablers (KASE) of Brazil, to build an AI Sovereignty Stack

**Luca Belli, Professor and Coordinator, Centre for Technology and Society at FGV Law School.**

## Abstract

An increasing number of countries is developing national strategies and regulatory proposals aimed at framing the use of artificial intelligence (AI). The main goal of this paper is to emphasise that the regulation of AI is only one of the essential elements that need to be considered to achieve AI Sovereignty. The paper defines AI Sovereignty as the capacity of a given country to understand, muster and develop AI systems, while retaining control, agency and, ultimately, self-determination over such systems. In this perspective, the paper proposes a layered framework, called "AI Sovereignty Stack" to analyse which elements are essential to achieve AI sovereignty. These all-important elements, which must be seen as interconnected and interdependent, are defined as "Key AI Sovereignty Enablers" or "KASE". Subsequently, the paper applies the proposed KASE to Brazil, to investigate whether the policy choices and governance arrangements of the tropical giant can allow it to assert AI Sovereignty or rather lead to a situation of AI dependency. The paper concludes by emphasising that the lack of AI Sovereignty is a situation shared by most countries and is particularly evident in the Global South, and argues that national governments should strive to revert AI dependency and build AI Sovereignty, to avoid a scenario of digital colonialism.[1]

[1] This paper is an extended version of an essay presented at the Digital Democracy Network Conference 2023, organised by the Carnegie Endowment for International Peace, and published in the conference collection, edited by Steven Feldstein. The author would like to thank Steven Feldstein and the participants of the Carnegie Endowment for International Peace's Digital Democracy Network Conference 2023 for their valuable feedback to an earlier version of this paper presented at the Conference.

## 2.1 Introduction

As a transformational technology,[2] Artificial Intelligence (AI) will have a global impact and considerable ramifications for national economies, democracies, and societies. While many countries are developing AI governance frameworks,[3] the main goal of this paper is to emphasise that the regulation of AI is only one of the essential elements that need to be considered to achieve AI Sovereignty.

AI Sovereignty is not a universally defined concept. In this paper, I put forward a definition of this concept, building upon what I have previously described as "Good Digital Sovereignty,"[4] thus considering AI Sovereignty as the capacity of a given country to understand, develop and regulate AI systems. I argue that AI Sovereignty should be seen as essential to retain control, agency, and self-determination[5] over AI systems.

[2] Jarvenpaa, S. L., & Ives, B. (1996). Introducing transformational information technologies: the case of the World Wide Web technology. *International Journal of Electronic Commerce, 1*(1), 95-126. <https://www.jstor.org/stable/27750802>.

[3] Belli, L., Curzi, Y., & Gaspar, W. B. (2023). AI regulation in Brazil: Advancements, flows, and need to learn from the data protection experience. *Computer Law & Security Review*, *48*, 105767. www

[4] Belli L. (June 2023). Building Good Digital Sovereignty through Digital Public Infrastructures and Digital Commons in India and Brazil. G20's Think20 (T20). <https://t20ind.org/research/building-good-digital-sovereignty-through-digital-public-infrastructures/> ; Belli, L. and Jiang, M. (Eds.). (Forthcoming). Digital Sovereignty from the BRICS Countries. Cambridge University Press.

[5] The right to self-determination is so-called a primary principle or principle of principles, as it plays an instrumental role to allow individuals to enjoy their human rights, thus being an enabler of other fundamental rights. For this reason, it is enshrined as the first article of both the Charter of the United Nations and the International Covenants of Human Rights. According to these three international-law instruments, states have agreed that "all peoples have a right to self-determination" and that "by virtue of that right they are free to determine their political status and to pursue their economic, social and cultural development." It is essential to emphasise the relevance of the internal dimension of self-determination, i.e. the right of peoples to freely determine and pursue one's economic, social and cultural development, including by independently choosing, developing and adopting digital technologies. Such conception is also corroborated by the recognition of the fundamental right to "informational self-determination" as an expression of the human right to have and develop a personality, first recognised by the German Supreme Court, in the 1983 Census case. The fundamental right to free development of personality is formally recognised internationally. Article 22 of the Universal Declaration of Human Rights affirms that "everyone is entitled to the realisation of the rights needed for one's dignity and the free development of their personality," while the International Covenant on Economic, Social and Cultural Rights consecrates this fundamental principle regarding the right of everyone to education and to participate in public life. Particularly, the Covenant's signatories have agreed that the right to education "shall be directed to the full development of the human personality and the sense of its dignity [.] and enable all persons to participate effectively in society" (Article 13.1). Moreover, the free development of personality is explicitly considered as instrumental to exercise the fundamental right "to take part in cultural life [and] to enjoy the benefits of scientific progress and its applications" (Article 15). See Belli, Luca. Network Self-Determination and the Positive Externalities of Community Networks. In L. Belli (Ed.) Community Networks: The Internet by the People for the People: Official Outcome of the UN IGF Dynamic Coalition on Community Connectivity. FGV. (2017: 35-64) <https://www.intgovforum.org/en/

In this perspective, I propose a layered framework to analyse which elements are essential to establish a country's AI sovereignty, defining them as "Key AI Sovereignty Enablers" or "KASE". Subsequently, I will analyse the case of Brazil, using the proposed KASE framework, to understand whether Brazilian policy choices and governance arrangements can allow the country to assert AI Sovereignty or rather lead to AI dependency.

I argue that sound governance,[6] regulation, research, and development in all the elements of the AI value chain are essential not only to achieve economic growth, social justice, and industrial leadership but, primarily, to assert (AI) sovereignty, avoiding the implementation of exclusively foreign AI systems in a country, which would likely transform the recipient country into a digital colony. Importantly, the purpose of this paper is not to advocate for AI autarchy, nor to deny the ample range of benefits that digital trade and cooperation can produce, but rather to discuss how countries could achieve a sufficient level of strategic autonomy, diversifying their AI value chains, and being able to grasp the functioning of AI systems, develop such systems rather than being mere consumers, and regulate them effectively.

The paper also emphasises that the careful consideration of each of the KASE and the importance of their interconnection, through an integrated approach, may allow countries to build what I define as an "AI Sovereignty Stack". This layered structure may reduce the country's exposure to the technological choices of foreign (private or public) actors, and simultaneously increase their agency and self-determination over and through AI systems.

filedepot_download/4391/1132> ; Belli, Luca *et al.* Cibersegurança: uma visão sistêmica rumo a uma proposta de marco regulatório para um Brasil digitalmente soberano. FGV Direito Rio. (2023: 69-94). <https://bibliotecadigital.fgv.br/dspace/handle/10438/33784>.

[6] For the purposes of this paper, governance is intended as the set of processes and institutional mechanisms that stimulate facilitate and organise coordinate the stakeholder interactions of different stakeholders in a political space, to confront different opinions and interests regarding a specific issue and, ideally, achieve the proposal of the best possible regulatory solution to frame such issues. Regulation is intended as the product of governance, consisting of an ample range of instruments that can foster the stability and proper functioning of complex systems, where the presence of multiple actors with varying or divergent interests can naturally lead to instability and dysfunction. Belli, Luca. De la gouvernance à la régulation de l'Internet. Paris: Berger-Levrault. (2016 :17-132).

Such interconnection must be reflected in the necessary coordination of research and development, governance and regulation of the various KASE to be able to form a well-functioning AI Sovereignty Stack. Such stack should be organised through a dedicated governance system allowing the authorities in charge of overseeing each KASE to cooperate with other authorities from different sectors (including with regulators of transversal sectors such as competition, consumer protection, data privacy, financial services, energy, and telecom infrastructure) to facilitate smooth organisation and, particularly, information sharing.

Importantly, this paper intends to adopt a pragmatic stance, stressing that achieving AI Sovereignty will be far from trivial, especially for Global South countries. However, in the perspective of the author, AI Sovereignty should be considered at least a policy priority. The KASE discussed in the next section require considerable planning, resources, and implementation capacity, but they should be – ideally – seen as a highly strategic objectives for the reinforcement of national sovereignty, allowing to resist possible adverse conditions, spanning from extraterritorial effects of foreign regulation, to the imposition of foreign sanctions and the increasingly frequent disruption of supply chains.

## 2.2 Presenting the Key AI Sovereignty Enablers (KASE)

In this paper I posit that the achievement of AI Sovereignty relies on the adoption of a systemic approach to AI, understanding the relevance and the interconnectedness of the Key AI Sovereignty Enablers (KASE). These elements are instrumental for ensuring that a country can develop, regulate, and utilise AI systems according to its own national interests, values, and strategic objectives, rather than being subject to the unavoidable impact of other (state or corporate[7]) entities' exercise of AI Sovereignty.

Importantly, AI Sovereignty is likely to become an increasingly relevant and strategic topic as the development and adoption of

[7] Luca Belli. Structural Power as a Critical Element of Digital Platforms' Private Sovereignty. In Edoardo Celeste, Amélie Heldt and Clara Iglesias Keller (Eds). *Constitutionalising Social Media*. (Hart 2022) <https://lucabelli.net/2021/08/10/structural-power-as-a-critical-element-of-social-media-platforms-private-sovereignty/>.

AI technologies continue to advance, acquiring a significant role in various aspects of society and democratic governance, not limited to the (digital) economy. The impact of AI advancement, which has been already the object of considerable research, especially concerning its interaction with data governance,[8] includes a wide range of critical sectors such as defence, infrastructural management, healthcare, and justice.

It seems important to emphasise that the capacity to develop and muster AI technology, rather than being regulated through it, does not rely exclusively on the elaboration and enforcement of well-crafted AI legislation. On the contrary, the achievement of an AI Sovereignty Stack entails the capacity to control and exercise agency and self-determination regarding at least eight different KASE that, together, compose the IA Sovereignty Stack, allowing the building of a sustainable and strategically autonomous AI ecosystem.

The fundamental elements that I define as KASE include sound (personal) data governance and algorithmic governance, strong computational capacity, meaningful connectivity, reliable electrical power, a digitally literate population, solid cybersecurity, and last, but not least, an appropriate regulatory framework. The next section analyses them, in the context of Brazil.

## 2.3 Exploring the KASE of Brazil

In this section, I will briefly present the KASE that compose what I define as the AI Sovereignty Stack, analysing how Brazil is harnessing each of them.

### 2.3.1 Data Governance

Data is the lifeblood of AI systems. Access to diverse, high-quality data is essential for training and improving AI models. Importantly, depending on the type of AI at stake, the data utilised to feed AI systems can be personal, governmental, confidential, copyrighted, etc, thus including a fair amount of complexity and need for regulatory compliance in the context of their processing. Hence, not only the

[8] CPDP LatAm. (2023, July 18). Publications – CPDP LaTAM 2023. CPDP LatAm 2023. <https://cpdp.lat/en/publications/>.

availability of large volumes of heterogeneous data is essential to develop AI capabilities, but having control over such data, including how they are collected, stored, processed, and transferred to third countries is a critical aspect of AI sovereignty.

Countries with large and diverse populations together with consolidated data collection practices and well-structured data policies will indubitably have a competitive advantage, constructing their AI sovereignty. It is important to emphasise that few countries enjoy the privilege of having both large data pools and sound data policies at their disposal. In this context, countries should consider establishing shared data policy frameworks, at regional level or within existing international governance mechanisms,[9] so that national data assets can be shared under substantially equal norms. This strategy would allow usage of much larger and diversified data pools, providing at the same time juridical certainty for AI researchers and developers, while protecting the rights of personal data subjects, intellectual property right holders, and preserving the public interest.

Particularly, sound data governance allows a country to protect its citizens' data privacy, ensure national and informational security, and harness the value of data for national development. Brazil made considerable progress in terms of data governance, by structuring one of the most progressive and refined open data policies[10] and by adopting a last-generation data protection framework, the *Lei Geral de Proteção de Dados* or LGPD.[11] The enforcement of the

[9] The finest example of international cooperation regarding data policy are provided by European initiatives. The Council of Europe Convention 108 is the most renown instance – and until the recent entry in force of the Malabo Convention, the only one – of international treaty regarding personal data protection. The most refined example of coordinated approach to data policy is offered by the European Union data policy framework, spanning form the General Data Protection Regulation, the Open Data Directive, and the most recent Data Act. It is important to stress that a less ambitious, yet relevant framework could also be proposed at the Latin American level, where most countries have already adopted similar data protection laws. In this regard, see Luca Belli, Ana Brian Nougrères, Jonathan Mendoza Iserte, Pablo A. Palazzi and Nelson Remolina Angarita. Hacia un modelo latinoamericano de adecuación para la transferencia internacional de datos personales. Centro de Tecnologia y Sociedad de Universidad de San Andrés. (2023).

[10] De Magalhães Santos, L. G., & Dhaou, S. B. Open Data and Emerging Technologies: Connecting SDG Performance and Digital Transformation. <https://cyberbrics.info/open-data-and-emerging-technologies-connecting-sdg-performance-and-digital-transformation/>.

[11] The Brazilian General Data Protection Law (LGPD) – Unofficial English Version <https://cyberbrics.info/brazilian-general-data-protection-law-lgpd-unofficial-english-version/>.

LGPD, however, remains still very embryonic, especially as regards new generative AI systems.[12]

Furthermore, personal data collection is considerably concentrated in the hands of a few foreign tech giants, primarily as a result of so-called zero-rating mobile Internet plans,[13] as discussed in the connectivity section below, thus frustrating the possibility to harness personal data as a national asset. Lastly, data security remains also very patchy[14] in the lack of a Cybersecurity law and given the lack of regulation on personal data security.

## 2.4 Algorithmic governance

Software algorithms are the foundation of AI systems, enabling them to perform tasks and make decisions. Importantly, algorithms can be the subject matter of regulation, but they can also play an instrumental role to elaborate regulation. On the one hand, the development and deployment of algorithms can – at least partly – give rise to risks and social problems triggering the need for regulatory intervention. On the other hand, algorithms can support the regulatory intervention itself, as they are increasingly useful and used to assist both the elaboration and implementation of regulation.

In this perspective, the development, deployment and regulation of or through algorithms are all equally important dimensions of algorithmic governance. Developing and owning proprietary software provides a considerable competitive advantage and allows for embedding normative values according to national specificities. Investing in research and development of AI algorithms, while also addressing the potential risks that they pose, can enormously enhance a country's technological capabilities, and reinforce AI Sovereignty.

Hence, the promotion of multistakeholder cooperation to develop software algorithms can allow for enhancing AI Sovereignty either when domestic players are stimulated to develop proprietary software,

[12] Belli, Luca. (2023, July 20). Why ChatGPT does not comply with the Brazilian Data Protection Law and why I petitioned the Regulator. *MediaNama*. <https://www.medianama.com/2023/05/223-chatgpt-brazilian-data-protection-law-ai-regulation/>.

[13] See <http://www.zerorating.info/>.

[14] Belli, L. (2021). The largest personal data leakage in Brazilian history. openDemocracy. <https://www.opendemocracy.net/en/largest-personal-data-leakage-brazilian-history/>.

or when software is developed in open-source through a collaborative process embraced – or even led – by national stakeholders. In this latter perspective, the first Lula Administration was a true pioneer in terms of a collective approach to digital sovereignty,[15] promoting free and open software (FOSS) as a strategic objective for national development, already in 2003. Such policy allowed not only to be strategically autonomous from foreign software producers but also to increase national understanding and development of software. Unfortunately, this policy was reversed by the Temer administration in 2016, de facto unleashing the recent phenomenon of platformisation of the public administration primarily through the use of foreign software providers.

Despite political turbulence, over the past two decades, Brazil has developed several industrial policy instruments aimed at fostering the national software industry. However, the software development sector has not become as thriving as it could, primarily due to a lack of consistency in software-related policies and the absence of policies focused on stimulating software development and implementation in an organic fashion, including by facilitating access to capital to jumpstart the domestic algorithm industry. Particularly, Brazilian software policies have lacked complementary instruments able to stimulate demand and supply, for instance through public procurements of nationally developed software, as happens commonly in China, or through the establishment of digital public infrastructures, as India did with the India Stack,[16] or by organising capacity building efforts aimed at fostering demand, as South Korea did in the late 1990s.

## 2.5 Computational Capacity

It is well-known that AI can require substantial computational resources for tasks such as training complex models and processing large datasets. Particularly, the most recent AI systems, such as generative AI, can be remarkably computer-intensive due to their increased

[15] Belli, L. (2023, March 1). Brasil precisa reconstruir sua soberania digital. Estadão. <https://www.estadao.com.br/politica/blog-do-fausto-macedo/brasil-precisa-reconstruir-sua-soberania-digital/>.

[16] See <https://indiastack.org/>.

complexity. Ensuring the existence or continuous access to sufficient computational capacity should be seen as a key strategic priority.

The availability of high-performance computing infrastructure depends on multiple factors, spanning from the accessibility of semiconductors and chips specifically designed for AI applications and last-generation Graphics Processing Units or GPUs, which are becoming particularly relevant to support (generative) AI, to specialised servers tailored to AI specificities that go into data centres. In this respect, it is interesting to note that some of the first policies adopted by the Lula 3 administration have been the reintroduction of the national support programme for the development of semiconductors (known as "PADIS", in its Portuguese acronym) as well as the suspension of the previous Bolsonaro administration decision to sell the National Center for Advanced Electronic Technology (Ceitec), which is the only semiconductors producer of Latin America.[17]

Moreover, it is essential to emphasise that the availability of cloud computing resources by itself is not enough to assert AI Sovereignty, which demands that cloud resources be not only available but fully compliant with national legislation. A telling example of how this is far from being the rule is offered by the online education platforms[18] provided by two major US tech companies in Brazil, which are supplied nationally and do not even mention how they comply with the Brazilian LGPD, despite the law being fully in force since 2021.

## 2.6 Meaningful connectivity

Meaningful connectivity, allowing users to enjoy reliable, well-performing, universally accessible Internet infrastructure for an affordable price plays an instrumental role for AI systems to function optimally and be used by the largest possible portion of the population. Seamless connectivity facilitates data exchange, collaboration, and access to cloud-based AI services. It enables real-time applications and supports the development and deployment of AI technologies

[17] Decree No. 11,456, of March 28, 2023. Amends Decree No. 10,615, of January 29, 2021, which provides for the Support Program for Technological Development of the Semiconductor Industry. <https://www.in.gov.br/en/web/dou/-/decreto-n-11.456-de-28-de-marco-de-2023-473390191>.

[18] Pacotes "education" do Google e da Microsoft não contemplam lei brasileira de proteção de dados. (n.d.). <https://aberta.org.br/pacotes-education-nao-contemplam-lgpd/>.

across various sectors, contributing to the construction of a country's AI Sovereignty.

Over the past ten years, Brazil has made enormous progress in terms of Internet penetration.[19] The cost of connectivity has considerably declined while the connected population has doubled in a decade. Yet, such a rosy picture hides less visible digital divides, which do not impinge on the quantity of but rather on the quality of Internet access. Most of the Brazilian "connected" population is considered so, but de facto only partially connected.

Indeed, more than 70% of the Brazilian connected population, and around 85% of the lower income population, has access primarily to a reduced set of apps included in so-called zero-rating plans,[20] based on not counting the data consumption of a few applications selected by the mobile internet operators. As such user attention ad user data collection is concentrated in a remarkably limited number of services, which typically are dominant social media platforms, thus making it particularly challenging for any other business to develop complete personal data sets that can be used to train AI models.

## 2.7 Reliable electrical power

As AI systems grow in relevance and size, they require a stable and increasingly relevant supply of electrical power[21] to operate effectively. Ensuring reliable power infrastructure and access to affordable electricity is necessary for maintaining uninterrupted AI operations. In this regard, it may be said that Brazil is probably one of the best-placed countries to support the expansion of AI infrastructure, as it is not only independent in energetic terms, but between 70% and 80% of its annual energy needs are satisfied via renewables, especially hydropower.

However, the national power grid is not exempted from criticism. In the short term, Brazil does not run the risk of a lack of energy

[19] TIC domicílios. (n.d.). Cetic.br – Centro Regional Para O Desenvolvimento Da Sociedade Da Informação. <https://cetic.br/pt/pesquisa/domicilios/publicacoes/>.

[20] IDEC (2021). Barreiras e limitações no acesso à internet e hábitos de uso e navegação na rede nas classes C, D e E. <https://idec.org.br/sites/default/files/pesquisa_locomotiva_relatorio.pdf>.

[21] Luccioni, S. (2023, April 12). The mounting human and environmental costs of generative AI. Ars Technica. <https://arstechnica.com/gadgets/2023/04/generative-ai-is-cool-but-lets-not-forget-its-human-and-environmental-costs/>.

supply thanks to the complementarity of various energy sources to hydropower, but the lack of structural planning and the possibility of adverse hydrology – which has been observed in recent years – can alter the cost of energy making it considerably higher. Hence, despite having developed a strong power infrastructure, the Brazilian capability to support the deployment of power angry technologies requires a stronger focus on planning to prevent potential dependency on external sources.

## 2.8 Digitally literate population

Enhancing the digital literacy of the population, through capacity building, training, and multigenerational education is essential not only to achieve a skilled AI workforce, but also to foster cybersecurity and, ultimately, national sovereignty.[22] Investing in AI education, research and development helps nurture a pool of talented AI professionals, while spreading an understanding of how to make the best use of technology. A sound educational strategy is therefore vital to allow the national population to gradually evolve from one being made primarily of consumers of digital technology into one composed of prosumers, i.e. individuals that can develop technology and produce innovation rather than being exclusively consumers.

Building a robust talent pipeline of AI researchers, engineers, and data scientists enables a country to develop and maintain its AI capabilities, increasing the possibility of being an exporter of technology and reducing the likelihood of becoming a digital colony. It is highly promising that the recently elected federal government has already adopted a new National Policy for Digital Education.[23]

However, it is still problematic to note that digital literacy keeps on being considered a priority only for the new generations of students, forgetting that literally no one in Brazil – as in most other countries – has received this type of education, thus remaining digitally illiterate. Such a situation is particularly risky in a context

[22] CyberBRICS. (2023, February 24). Cybersecurity and digital sovereignty: a new path for Brazil. CyberBRICS. <https://cyberbrics.info/cybersecurity-and-digital-sovereignty-a-new-path-for-brazil/>.

[23] Law No. 14.533 – Brazil, Jan. 11, 2023. <https://www.planalto.gov.br/ccivil_03/_ato2023-2026/2023/lei/L14533.htm.

of accelerated digital transformation and automatization, in which understanding the functioning of technology becomes a primary necessity not only for the youngest generation but especially for all the individuals, whose labour, social and economic conditions are likely to be affected by the deployment of AI systems.

## 2.9 Strong cybersecurity

AI systems are susceptible to cybersecurity threats and can be used to perpetrate cyberattacks. Robust cybersecurity measures are vital for any country but become even more so in the context of increasingly accelerated digital transformation and deployment of AI systems. Particularly, protecting AI critical infrastructure from cyberattacks is essential. Brazil has recently enacted a considerable number of sectoral cybersecurity regulations,[24] spanning the telecom sector, the banking sector, the electricity sector, and the personal data protection laws. While much progress has allowed the country to climb the International Telecommunications Union's Cybersecurity Index,[25] it must be noted that this positive advancement must be considered again with a grain of salt.

Indeed, Brazil still lacks a Cybersecurity Law and a National Cybersecurity Agency, although they have been recently proposed by a study produced by the Center for Technology and Society at FGV[26] and by a Draft Bill formulated by the Brazilian Presidency.[27] The existence of a highly fragmented approach to cybersecurity, driven by the initiatives of sectorial agencies with no general competence in cybersecurity, and frustrated by the lack of coherent national strategies on cybersecurity is probably one of the main vulnerabilities of the countries, which have not yet managed to create a solid governance framework to connect, coordinate, and

[24] Belli, L. *et al.* (2023). Cibersegurança: uma visão sistêmica rumo a uma Proposta de Marco Regulatório para um Brasil Digitalmente soberano. CyberBRICS. <https://cyberbrics.info/ciberseguranca-uma-visao-sistemica-rumo-a-uma-proposta-de-marco-regulatorio-para-um-brasil-digitalmente-soberano/>.

[25] Brazil rises in international cybersecurity ranking. (2022, June 24). Serviços E Informações Do Brasil. <https://www.gov.br/en/government-of-brazil/latest-news/2022/brazil-rises-in-international-cybersecurity-ranking>.

[26] Belli, L et al. (2023).

[27] PNCiber Draft Bill <https://www.gov.br/gsi/pt-br/composicao/SSIC/dsic/audiencia-publica/PNCiberAudienciaPublicaProjetoBase.pdf>.

leverage the incredible amount of talent that Brazil produces in terms of cybersecurity.

## 2.10 Appropriate regulatory framework

A comprehensive governance framework that encompasses ethical considerations, data protection laws, and AI regulations is crucial for AI sovereignty. Establishing clear guidelines and standards for AI development, deployment, and usage ensures responsible and accountable AI practices. In this perspective, the Brazilian Congress is discussing a new Bill for an AI Regulatory Framework[28] to help protect citizens' rights, promote fairness, and prevent discrimination and other potential risks, thus aiming at steering the development, deployment, and use of AI technologies sustainably.

It is important to note that, while this initiative is surely laudable, even if still ongoing, it is not yet clear to what extent it will be able to effectively address the regulation of AI. The latest version of the proposed Bill includes many terms which provide a necessary level of flexibility on key issues such as AI systems transparency, data security, data governance or risk management. However, such flexibility, which is welcome to craft a law that can adapt to technological evolution, must be matched with a mechanism that allows the specification through regulation or standardisation.

In the absence of such specifications, the law risks being highly ineffective. In this regard, it is necessary to consider the recent Brazilian experience regulating data protection to understand that the adoption of modern law and the establishment of a new regulatory authority is only the beginning of the regulatory journey, which risks being considerably jeopardised when the enormously pressing task of specifying the law is attributed to a regulator that seems to be purposefully created being "ineffective by design".[29]

[28] PL 2338/2023 – Senado Federal. (s.d.). <https://www25.senado.leg.br/web/atividade/materias/-/materia/157233>.

[29] New Data Architectures in Brazil, China, and India: From Copycats to Innovators, towards a post-Western Model of Data Governance Luca Belli. (n.d.). IJLT. <https://www.ijlt.in/journal/new-data-architectures-in-brazil%2C-china%2C-and-india%3A-from-copycats-to-innovators%2C-towards-a-post-western-model-of-data-governance>.

## 2.11 Conclusions

It is important to reiterate that the abovementioned AI Sovereignty enablers are interconnected and mutually reinforcing. This consideration is particularly relevant in a moment where legislators and governments around the world are studying the regulation of AI, frequently ignoring the utmost importance of the other fundamental elements that I define as KASE. Considering the interconnectedness of the KASE and leveraging their interdependence through an integrated approach is essential to achieve AI Sovereignty and avoiding digital colonialism.

However, such an approach seems to be absent from the current Brazilian "strategic" vision for AI. Indeed, anyone analysing the 2021 Brazilian Artificial Intelligence Strategy (EBIA)[30] will immediately notice the lack of strategic elements in the strategy. The document has been the object of unanimous critiques from observers as it merely includes general considerations about how AI could be implemented in several sectors, without defining neither the elements that may allow coordinating the implementation of the strategy, nor those that can allow assessing such an implementation, or who would be responsible for such implementation.

By providing a preliminary understanding on what are the essential elements that countries need to consider in their strategic approach to AI, this paper also aims at offering some food for thought that could inspire the revision of the Brazilian strategic approach to AI by the current administration. As noted, an integrated approach considering the KASE is instrumental to achieve AI Sovereignty, developing indigenous AI capabilities, diversifying supply chains, increasing the digital literacy of the population, fostering strategic investments and partnerships, and safeguarding the security of critical AI infrastructure.

It is important to be realistic and acknowledge that not all countries might be able to elaborate and implement the necessary strategic, policy and institutional changes allowing them to build an AI Sovereignty Stack. Such an effort might be especially herculean

[30] Gaspar, W. (2022, March 28). Artificial Intelligence in Brazil still needs a strategy. CyberBRICS. <https://cyberbrics.info/artificial-intelligence-in-brazil-still-needs-a-strategy/>.

for Global South countries, which typically depend on foreign technologies. However, a careful mix of creative thinking and – much needed – political vision regarding technological development may allow to overcome some of the most burdensome obstacles for low-income countries, for instance by embracing the use of open software to overcome the considerable financial costs determined by dependency on foreign software. The elaboration of an AI Sovereignty Stack, therefore, should be seen as an ideal goal that all countries should strive to achieve but that may not be feasible for all countries.

Ultimately, countries that possess strong capabilities in the KASE areas are not only better positioned to maintain control over their AI technologies, policies, and data, but they will likely increase their technological relevance, reducing dependence on external sources and preserving their national interests and autonomy in the AI landscape. Countries lacking such capability need to reconsider thoroughly their strategic approaches to AI, to minimise the considerable risks prompted by AI dependency that the already ongoing phenomenon of digital colonialism[31] is likely to exacerbate.

[31] See e.g. Avila Pinto, R. (2018). Digital sovereignty or digital colonialism? New tensions of privacy, security and national policies. SUR: International Journal on Human Rights, 15(27), 15-27; Couldry, N. & Mejias, U. (2019). The costs of connection: How data Is colonizing human life and appropriating it for capitalism. Stanford, CA: Stanford University Press.

# 3 An Assessment of the Key AI Sovereignty Enablers within the South African context

**Melody Musoni, Policy Officer, European Centre for Development Policy Management, The Netherlands;**

**Sizwe Snail ka Mtuze, Attorney of the South African High Court, Adjunct Professor, Nelson Mandela University, Visiting Professor, Center for Technology & Society (CTS) at FGV Law School, Rio de Janeiro.**

## 3.1 Abstract

African countries are taking steps in carving out their position as competitors in the development of Artificial Intelligence (AI). The mantra ‘AI made in Africa for Africa’ is guiding some Africans to take decisive actions to strategically position themselves as AI sovereigns with the authority to create enabling environments which promote local innovations and use of AI tools to develop home-grown solutions, control all actors in the AI and data markets, and equipping citizens with the requisite AI digital skills. This paper assesses the Key AI Sovereignty Enablers (KASE) framework proposed by Belli within the South African context. The paper provides recommendations on the way forward reading KASE in South Africa.

## 3.2 Introduction

African countries are taking steps in carving out their position as competitors in the development of Artificial Intelligence (AI). The mantra ‘AI made in Africa for Africa’[32] is guiding some Africans to take decisive actions to strategically position themselves as AI sovereigns with the authority to create enabling environments which promote local innovations and use of AI tools to develop home-grown solutions, control all actors in the AI and data markets, and equipping citizens with the requisite AI digital skills. Leading the pact of the Smart Africa Alliance[33] is South Africa which is running the country flagship on AI with a clear mandate of strengthening

[32] Smart Africa Blueprint: Artificial Intelligence, 2021. <https://smart.africa/board/login/uploads/70029-eng_ai-for-africa-blueprint.pdf>.

[33] Smart Africa Alliance is the coming together of African Heads of State and Government with a shared interest in driving socio-economic development on the continent using technology. Smart Africa member states each lead a specific project related to digital <https://smartafrica.org/who-we-are/>.

local technical know-how on AI, removing entry barriers to AI and developing policy frameworks ready for AI.

South Africa currently lacks legislation, regulation or official policy that dictates or guides the ethical use of AI.[34] Very little legal literature has been written therein. Adams sums up other definitions of foreign authors and defines AI as "… the simulation of human intelligence by allorhythmia, computer programmes and machines …".[35] Gravett citing the works of Turing[36] as well as, Shubendu and Vijay[37] defines same as…:

> *"…[A] computer's ability to imitate human intelligent behaviour, especially human cognitive functions, such as the ability to reason, discover meaning, generalise and learn from past experience … machines that respond to stimulation consistent with traditional responses from human, given the human capacity for contemplation, judgment and intention."*[38]

Through the leadership of President Ramaphosa, the Presidential Commission on the 4th Industrial Revolution (4IR) was established to guide the country in the development of a strategy on 4IR (including AI).[39] The Presidential Commission on 4IR produced a report (hereafter "PC4IR Report") [40] which came up with 8 (eight) key recommendations which are the establishment of an AI Institute; investment in human capital; improving the industrial policies on manufacturing and new materials; secure and avail data to enable innovation; incentivise future industries, platforms and applications for 4IR technologies; building 4IR infrastructure; reviewing, amendment

[34] Adams, NR. South African Company Law in the Fourth Industrial Revolution: Does Artificial Intelligence create a need for legal reform? LLM Thesis- Wits (2021) p, 13 (hereafter 'Adams, 2021').

[35] *Ibid.*

[36] Turing, A. Mind (1950) in *Computing Machinery and Intelligence* 59(236), 4337.

[37] Shubhendu, S and Vijay, J. Applicability of Artificial Intelligence in Different Fields of Life. (2013) *in International Journal of Scientific Engineering and Research* Volume 1 Issue 1.

[38] Gravett, WH. Is the Dawn of the Robot Lawyer upon us? The Fourth Industrial Revolution and the Future of Lawyers. in *PER* (2020), Volume 23. <http://www.scielo.org.za/scielo.php?script=sci_arttext&pid=S1727-37812020000100024&lng=en&nrm=iso> (hereafter 'Gravett,2020,*PER*').

[39] Presidential Commission on 4IR. Government No. Gazette 42388, 9 April 2019 <https://www.gov.za/documents/presidential-commission-fourth-industrial-revolution-members-and-terms-reference-9-apr>.

[40] Presidential Commission on Fourth Industrial Revolution Report. Government Gazette No. 43834, 23 October 2020. <https://www.gov.za/sites/default/files/gcis_document/202010/43834gen591.pdf>.

or adoption of new policy and legislation; and establish 4IR strategy and implementation coordination council in the Presidency.[41]

The establishment of an AI Institute, development of an AI strategy, and implementation of the Master Plan have all been included in the Department of Communications and Digital technologies (DCDT) Annual Performance Plan (APP) 2022-23.[42] These initiatives dovetail with existing DCDT programmes, such as the Digital and Futures Skills Implementation Programme and associated AI Training Network.[43] South Africa is developing its capacity to be in control of the complete AI value chain from collection of AI datasets, analysis of data, development of AI solutions and services and controlling or having influence over all the actors involved in the AI lifecycle through regulation, digital industrial policies and strategic partnerships. South Africa, being a developing country with a history of inequality and discrimination, requires that AI be used as a strategic developmental tool. If not carefully managed and regulated, AI can be detrimental and has the potential to exacerbate the inequalities and discrimination.[44]

South Africa is characterised by a relatively dispersed digital policy and regulatory environment, with several public entities shaping the development and use of technology. Most of these entities are now considering the emerging impact of AI. It is therefore important that we start understanding whether the various policy initiatives are likely to be complementary, and the extent to which they may support a responsible approach to AI adoption.[45] Hlomani proposes that this requires that a context specific approach to regulation of AI and National AI strategy be observed in South Africa which is currently unregulated by law.[46]

[41] *Ibid.*

[42] DCDT Annual Performance Plan (APP) 2022-23 p.17. <https://www.dcdt.gov.za/documents/annual-performance-plans/file/207-annual-performance-plan-2022-2023.html>.

[43] UNESCO. Landscape study of AI policies and use in Southern Africa: research report. (2022). Conference Proceedings of Southern Africa sub-Regional Forum on Artificial Intelligence, Windhoek, p.35.

[44] Hlomani, H. Why South Africa needs a more holistic and contextualized approach AI regulation. *Daily Maverick* <https://www.dailymaverick.co.za/article/2023-05-23-why-south-africa-needs-a-more-holistic-and-contextual-approach-to-ai-regulation/>.

[45] UNESCO. Landscape study of AI policies and use in Southern Africa: research report. (2022). Conference Proceedings of Southern Africa sub-Regional Forum on Artificial Intelligence, Windhoek, p.35.

[46] Hlomani, H. Why South Africa needs a more holistic and contextualized approach AI regulation. *Daily Maverick* <https://www.dailymaverick.co.za/article/2023-05-23-why-south-africa-needs-a-more-holistic-and-contextual-approach-to-ai-regulation/>.

According to Belli, AI Sovereignty (hereafter "AIS") is not a universally defined concept.[47] Belli defines AIS as , "*the capacity of a given country to understand, muster and develop AI systems, while retaining control, agency, and self-determination over such systems*".[48] In this paper, we analyse the 8 point Key AI Sovereignty Enablers (KASE)[49] making up the AIS Stack and how the framework applies to the South African context. The key enablers making up the proposed KASE framework consists of data governance; algorithmic governance; computational capacity; meaningful connectivity; reliable electrical power; digitally literate population; strong cybersecurity; and appropriate regulatory framework. In our discussion, we also point out the critical areas that South Africa is prioritising to assert its AIS on the continent.

## 3.3 Key AI Sovereignty Enablers (KASE)

### 3.3.1 Data governance

Adams observes that the right to privacy, as provided for in section 14 of the Constitution,[50] affords South African citizens a level of privacy which extends beyond most privacy protection in Southern African States.[51] It is important to mention that the same is not absolute and subject to reasonable expectation of privacy.[52] The different adages 'data is the new oil' or 'data is the new gold' are analogies to signify the value of data (different types of data from personal, non-personal, government data, company data, proprietary data, open data) and its importance in transforming and powering digital economies. South Africa is determined to assert its AI sovereignty by regulating the whole data value chain.[53] Recent policy proposals in South Africa indicate the special interest for the government to

[47] Belli, L. Exploring the Key AI Sovereignty Enablers (KASE) of Brazil, towards an AI Sovereignty Stack. In Carnegie Endowment for International Peace. Digital Democracy Network Conference 2023 Essay Collection. (2023).

[48] *Ibid.*

[49] *Ibid.*

[50] Constitution of South Africa, Act 108 of 1996.

[51] Adams, R et al. (2021) Human Right and the Fourth Industrial Revolution in South Africa, p.27.

[52] See *Haze Club (Pty) Ltd and Others v Minister of Police and Others* (2101/2021) [2022] ZAWCHC 269; [2023] 1 All SA 280 (WCC) at par 39.

[53] The lack of a legal presence of some of the AI companies in South Africa makes it difficult for it to dictate how they should process data and exert its sovereignty over these companies. The proposed policy interventions are still vague and unclear on how South Africa will successfully regulate data produced by the private sector.

extend and exert control over the production and storage of data which includes gaining control over the big tech dominating the data value chain as well as improving data flows between government institutions for planning purposes. If approved, the National Data and Cloud Policy[54] (hereafter "NDCP")[55] would permit citizens to access government data for economic purposes while the government will also have access to critical data held by the private sector to enhance government planning and improve service delivery.[56] Strategically placing the government in a position to gain access and control over data from the private sector shines a glimmer of hope that South Africa will assert its AI sovereignty and not be a victim of digital colonialism,[57] a worrisome circumstance that many African countries find themselves in.

Personal data and use of AI in automated decision making processes and profiling are carefully regulated under the Protection of Personal Information Act (hereafter "POPIA") which gives effect to section 14 of the Constitution.[58] Earlier fears on enforcement of POPIA have been allayed as the Information Regulator has been quite astute to exercising its powers in politically charged cases[59] and those against big pharma.[60] It remains to be seen how the Information Regulator (hereafter "IR") will handle some of the uses of AI, especially

[54] National Data and Cloud Policy Government Gazette 44389 Government Notice 306, 1 April 2021 <https://www.gov.za/sites/default/files/gcis_document/202104/44389gon206.pdf>. Earlier in April 2023, the government announced that it was finalizing the final version of the NDCP. <https://www.sanews.gov.za/south-africa/government-finalising-national-data-cloud-policy>. At the time of writing, it is still not clear when the final version will be published.

[55] Ibid.

[56] Ibid.

[57] Musoni et al argued that digital colonialism impacts national data governance frameworks. Musoni, M, Karkare, P, Teevan, C and Domingo, E. Global approaches to digital sovereignty: Competing definitions and contrasting policy. ECDPM Discussion Paper No. 344. <https://ecdpm.org/work/global-approaches-digital-sovereignty-competing-definitions-and-contrasting-policy>.

[58] POPIA is the principal data protection law in South Africa which provides a list of minimum requirements which must be met when processing personal data. Specific to AI are the provisions of section 71 of POPIA where it is prohibited to make decisions with legal consequences by relying entirely on automated decision making. The section also provides exclusions where automated processing and profiling is permissible.

[59] The Information Regulator issued an enforcement notice against the South African Police Service <https://inforegulator.org.za/wp-content/uploads/2020/07/ENFORCEMENT-NOTICE-SAPS-MATTER-04052363.pdf>; The Information Regulator also issued an enforcement notice against the Department of Justice and Constitutional Development where it is housed under <https://inforegulator.org.za/wp-content/uploads/2020/07/ENFORCEMENT-NOTICE-DOJCD-MATTER-090523.pdf>.

[60] <https://inforegulator.org.za/wp-content/uploads/2020/07/FINAL-MEDIA-STATEMENT-ENFORCEMENT-NOTICE-ISSUED-TO-DISCHEM-PHARMACIES-LTD.pdf>.

facial recognition software used in public spaces. The IR recently approved codes of conduct for the banking industry and credit bureaux which includes aspects on automated decision making and profiling. However, the codes of conduct are very descriptive and do not exclusively deal with AI which means the IR may need to develop guidance notes on use of AI tools and application with POPIA.

The IR may start by endorsing resolutions passed by the Global Privacy Assembly (hereafter "GPA") as this can guide the IR when developing guidance notes on use of AI in South Africa. Some of the notable resolutions include the 2020 GPA Adopted Resolution on Accountability in the development and use of Artificial Intelligence,[61] the 2020 GPA Resolution on Facial Recognition Technology[62] or the 2022 GPA Resolution on Principles and Expectations for the Appropriate Use of Personal Information in Facial Recognition Technology.[63] It is submitted that a strong data governance legal framework furthers the aims and objectives of AIS and that same be regularly reviewed and updated with the changing times.

## 3.4 Algorithmic governance

According to Olorunju, discussions on the Global South concerning Algorithmic Governance (hereafter 'AG') are often generalised and fail to consider the differential infrastructural, institutional and human rights concerns within the African Continent. Olorunju further states that there is insufficient research and data from other contexts which has resulted in drafting of misguided and ineffective policies that are of limited benefit to Africans.[64] There is a need to follow an African – Centred Approach[65] which includes incorporation of

[61] <https://globalprivacyassembly.org/wp-content/uploads/2020/11/GPA-Resolution-on-Accountability-in-the-Development-and-Use-of-AI-EN.pdf.

[62] <https://globalprivacyassembly.org/wp-content/uploads/2020/10/FINAL-GPA-Resolution-on-Facial-Recognition-Technology-EN.pdf.

[63] Resolution on Principles and Expectations for the Appropriate Use of Personal Information in Facial Recognition Technology. <https://globalprivacyassembly.org/wp-content/uploads/2022/11/15.1.c.Resolution-on-Principles-and-Expectations-for-the-Appropriate-Use-of-Personal-Information-in-Facial-Recognition-Technolog.pdf.

[64] Olorunju, N. African Algorithmic Governance: Benefit of a Community-based Approach. (2022) <https://researchictafrica.net/2022/04/03/african-algorithmic-governance-benefit-of-a-community-based-approach/.

[65] An African Centred Approach to AI is characterised by ethical principles and value-based approaches that arise from distinctly African histories and value systems, this includes incorporation of collective rights and active participation of communities historically marginalised from debates around AI and data governance. Ibid.

collective rights and community practices on data governance. Traditional African governance frameworks are based up community based approaches to governance. The need to consider indigenous knowledge systems and cultural norms cannot be underscored enough in the African context.[66]

For that reason, it is important to look at ethical principles and value-based approaches that arise from distinctly African histories and value systems to build locally relevant and appropriate policy and governance solutions. This is important when much of the discussion around AG to date has centred on a principle-based approach in the form of ethical principles and standards on AI largely developed in the Global North. South Africa's digital and industrial policies seem to advocate for the development of domestic technologies and software in order to meet local needs as well as developing solutions and products to export to other countries. The country is exploring different channels to promote software development and AI algorithms using local data, local entities and promoting the development of local skillsforce. Several initiatives from investment in innovation hubs, Public-Private Partnerships (PPPs), funding for Research and Development (R&D), have been launched to promote research and development of AI. For instance, an AI-based algorithm with the capability to detect COVID-19 cases was developed by the Gauteng Provincial Government in partnership with iThemba Labs, the University of Witwatersrand and the University of York, Canada.[67] Government entities are also encouraged to procure services of software developers to develop proprietary software which will be owned by the government.[68] By developing its own algorithms and software, South Africa will be better placed to train the algorithm on ethical considerations such as bias and discrimination and embed normative values[69] of ubuntu.

[66] *Ibid.*

[67] <https://www.wits.ac.za/news/latest-news/research-news/2021/2021-03/ai-powered-algorithm-released-to-detect-the-third-wave-in-south-africa.html>.

[68] Department of Public Service Administration. Policy on Free and Open Source Software Use for South African Government <https://www.gov.za/sites/default/files/gcis_document/201409/fosspolicy0.pdf>.

[69] Belli, L. Exploring the Key AI Sovereignty Enablers (KASE) of Brazil, towards an AI Sovereignty Stack. In Carnegie Endowment for International Peace. Digital Democracy Network Conference 2023 Essay Collection. (2023).

## 3.5 Computational Capacity

AI development will depend on the availability of computing infrastructures to host, process and use data to enable data analytics and machine learning.[70] Data centres play a significant role in providing the processing capacity, storage solutions and delivering an integrated AI infrastructure, applications and services. South Africa is preparing itself not only to have computational capacity for its own needs but also to be an attractive host to the data centre industry in the African continent.[71] The approach it anticipates aims at replacing the current actors on the data market (mainly foreign owned entities) with locally owned entities, changing the culture of doing business in the cloud market by insisting on local processing and local storage of data instead of use of overseas cloud data centres and defines the terms for data use and data sharing. This proposed approach has been criticised for being vague and sometimes using incorrect references to data-related concepts and over-emphasising minor benefits of data localisation.[72] It is anticipated that the revised NDCP will focus on better regulation of foreign-owned infrastructure instead of indigenising all existing infrastructure, while also encouraging public-private partnerships.

The proposed policy intervention is to establish a High-Performance Computing and Data Processing Centre (HPCDPC) which will include processing and data facilities and cloud computing capacity and will consolidate existing public funded data centres.[73] The HPCDPC shall provide use-on-demand cloud services for State entities, national departments, provinces, municipalities, metros, SOEs, universities, research centres, civil society organisations, and local businesses. The country has a window of opportunity to build on existing capacity and capability though the Centre for High Performance Computing[74]

[70] Smart Africa AI Blueprint.

[71] NDCP.

[72] Van der Berg, S. Data protection in South Africa: The potential impact of data localisation on South Africa's project of sustainable development. Mandela Institute Policy Brief 02, 2021. Razzano, G. Data localisation in South Africa: Missteps in the valuing of data. Mandela Institute Policy Brief 06, 2021. Research ICT Africa. 'Written submission in response to the: Proposed National Data and Cloud Policy'. 2021. <https://researchictafrica.net/wp/wp-content/uploads/2021/06/RIA_Submission_DATA_and_Cloud_Policy.pdf>.

[73] NDCP.

[74] <https://www.nicis.ac.za/chpc/>.

and other existing State compute resources.[75] The sustainability of the HPCDPC is strongly dependent on reliable power supply, internet connectivity and availability of resources to manage and sustain the computing systems, areas that South Africa is currently battling with.

## 3.6 Meaningful Connectivity

An efficient digital public infrastructure which enables people to connect to fast, reliable and affordable internet coupled with having access to data, interconnected ICT devices and interoperable systems are important prerequisites for people to use AI tools to innovate and for social and economic development. South Africa is doing fairly well in the Africa region in providing internet connectivity by investing in submarine fibre, last mile broadband connectivity and 4G and 5G spectrum roll out.[76] Of the 60.14 million people in South Africa, 43.48 million have been active internet users at the beginning of 2023, with 25.80 million of those people using social media.[77] This means over 16.66 million people in South Africa still do not have access to the internet due to both infrastructure challenges and affordability. Despite the challenges of broadband infrastructure, the number of internet users in South Africa has significantly increased, and progress of South Africa's digital ecosystem has not been completely hindered. This positive trend is owing to the fact that a major portion of South Africans are using their mobile devices to access the internet.[78] Despite improvements to make data affordable, such as introduction of low-cost data packages and zero-rated government websites and off-peak data packages, an estimated 42% of the population cannot still afford the internet due to earning below minimum wage.[79]

[75] PC4IR Report.

[76] <https://www.icasa.org.za/news/2023/icasa-publishes-three-final-radio-frequency-spectrum-assignment-plans-for-high-demand-spectrum>.

[77] <https://datareportal.com/reports/digital-2023-south-africa?rq=digital%202023%20%3A%20south%20africa>.

[78] PC4IR Report.

[79] Freedom House. South Africa: Freedom on the net. 2021 <https://freedomhouse.org/country/south-africa/freedom-net/2021>.

## 3.7 Reliable Electrical Power

Hyper scale data centers, the emergence of 5G and interconnected devices have led to higher energy consumption, with a third of all generated electricity predicted to be used only by data centres. The challenge for South Africa is that an increase in power consumption and the energy demands are putting a strain on its aging power infrastructure operated by Eskom.[80] This has resulted in country wide daily power cuts or loadshedding undermining the country's economic recovery from COVID-19.[81] Unreliable power supplies impose restrictions on innovation and reduce the number of people and amount of hours spent in using AI tools like generative AI to create content and develop solutions. The government is aware of the limitations on relying on coal-powered energy and efforts are being made to look into clean and green energy sources, leveraging AI to improve energy efficiency, greening data centers and relying on power supply from independent power producers[82] to reduce the total dependence on the strained national electricity grid.[83]

## 3.8 Digitally Literate Population

South Africa hopes to build its own pool of AI experts to research and develop AI driven solutions to address local problems. The mantra 'AI made in Africa for Africa' is only achievable if and when Africans are digitally skilled to become prosumers, entrepreneurs and innovators. South Africa's Digital and Future Skills Strategy[84] provides strategic points to enhance digital skills, through various programmes targeting the different literacy levels of users, their needs and the needs of different sectors. For South Africa to fully realise the benefits of a digital economy and AI, it should adopt an integrated skills development plan and programme, designed to ensure the building of competencies that will enable the majority of South Africans to

[80] PC4IR Report. Ongoing energy crisis is also thwarting economic growth. IMF 'South Africa's Economy Loses Momentum Amid Record Power Cuts'. <https://www.imf.org/en/News/Articles/2023/06/15/cf-south-africas-economy-loses-momentum-amid-record-power-cuts>.

[81] <https://www.investec.com/en_za/focus/economy/sa-s-load-shedding-how-the-sectors-are-being-affected.html>.

[82] South Africa's Industrial Policy Action Plan 2017/18 – 2019/20.

[83] NDCP.

[84] National Digital and Future Skills Strategy South Africa Government Gazette 43730, 30 August 2020 <https://www.gov.za/sites/default/files/gcis_document/202009/43730gen513.pdf>.

understand the fundamentals of AI, data and cloud computing, and how to access these to exploit economic opportunities.[85] In 2023, the Department of Basic Education introduced robotics and AI in primary schools as part of a pilot program with the target of implementing a full curriculum across all grades in 2024.[86] The challenge for South Africa will be the availability of technical, financial and human resources to address the digital illiteracy rate.

## 3.9 Strong Cybersecurity

AI development depends on secure, reliable and trustworthy data processing systems. Cyber attacks are on the rise and several data breaches and data leakages have dampened public confidence in IT systems. According to a report by Interpol, South Africa is leading the continent in the number of cybersecurity threats.[87] The National Cybersecurity Policy Framework[88] guides the implementation of cybersecurity initiatives and measures. the Cybercrimes Act[89] and the Critical Infrastructure Protection Act[90] are the core pieces of legislation regulating cybercrime in South Africa. Cybersecurity is also promoted by Section 19 of POPIA, according to which responsible parties must also apply appropriate and reasonable technical and organizational steps. Although security of personal information is usually associated with technical ICT measures, the security of physical records should not be ignored – as much as we accommodate electronic communications and records in our modern legal discourse, we should not forget the effect that organizational measures will have on documents as known in the bricks and mortar world.[91]

[85] NDCP Section 10.8.

[86] <https://www.education.gov.za/CodingRoboticsPilot.aspx>.

[87] The top cyberthreat trends in Africa relate to Business Email Compromise, phishing, ransomware attacks, banking trojans and stealers, online scams, cyber extortion and crime as a service. African Cybercrime Operations Desk. 'African cyberthreat assessment report: Cyberthreat trends outlook' Interpol 2023.

[88] National Cybersecurity Policy Framework. Government Gazette No. 39475 GN 609. 04 December 2015. <https://www.gov.za/sites/default/files/gcis_document/201512/39475gon609.pdf>.

[89] Cybercrimes Act 19 of 2020.

[90] Critical Infrastructure Protection Act 8 of 2019.

[91] De Stadler, E and Esselaar, P. A Practical Guide to the Protection of Personal Information Act (2015) p.35.

POPIA offers important guidance on how to reach an adequate degree of information security, which can be structured into four steps to be completed for compliance with Section 19(2) of the POPIA, namely:

- risk identification;
- establishment and maintenance of appropriate safeguards;
- verification of effective implementation; and
- updating safeguards.[92]

However, the country is yet to develop a National Cybersecurity Strategy and a specific law on Cybersecurity is yet to be promulgated.

## 3.10 Appropriate Regulatory Framework

One of the principal regulatory challenges confronting Africa relates to the fact that AI regulation interplays with a multiplicity of factors and elements such as how AI systems deal with multiple regions and industries, including intellectual property and civil liability challenges, data protection, cybersecurity and ethical considerations.[93] The African Union Commission on Human and People's Rights called for State Parties "to work towards a comprehensive legal and ethical governance framework for AI technologies, robotics and other new and emerging technologies so as to ensure compliance with the African Charter and other regional treaties."[94]

As previously stated herein, South Africa currently lacks legislation, regulation or official policy that dictates or guides the ethical use of AI.[95] Existing legislation may be loosely translated to apply to AI, but the application is both generic and limited. Further, there is no national AI strategy but piecemeal strategies entrenched within national plans and policies point towards shaping AI policy. The existing policies focus almost exclusively on economic development

[92] *Ibid.*

[93] Smart Africa AI Blueprint.

[94] African Commission on Human and Peoples' Rights. Resolution on the need to undertake a Study on human and peoples' rights and artificial intelligence (AI), robotics and other new and emerging technologies in Africa. Resolution 473. <https://achpr.au.int/en/adopted-resolutions/473-resolution-need-undertake-study-human-and-peoples-rights-and-art>.

[95] Adams N. South African company law in the fourth industrial revolution: does artificial intelligence create a need for legal reform? (LLM dissertation, University of the Western Cape, 2021). <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4052285>.

and not on the appropriate use or ethical issues associated with AI.[96] This makes it a priority for South Africa to develop an appropriate regulatory framework and policy strategy on AI. Musoni proposes that before adoption of AI specific regulatory frameworks, countries should prioritise strengthening the existing foundational frameworks on data governance and use AI regulatory sandboxes as testbeds.[97]

It is suggested that South Africa like many other African countries explore developing national AI Strategies to guide AI regulation adoption.[98] Adams, citing Dulton, also recommends that the said strategies and legislation follow global trends in AI regulation such as: basic and applied research in AI; AI talent attraction, development and retainment; future of work and skills; industrialisation of AI technologies; public sector use of AI; data and digital infrastructure, ethics [and human rights]; regulation; inclusion and foreign policy.[99] To ensure policy and legal interoperability, South Africa should consider the auspices of regional frameworks like AUC Resolution 473, and international frameworks like the UNESCO Recommendation on the Ethics of AI.[100] The adopted AI policies should also advance gender equality through digital literacy and the inclusion of more women in digital spaces.[101] South Africa can also take lessons from other African countries like Rwanda[102] and Egypt,[103] which have developed their own AI strategies and policies.

[96] Ormond, E. Global To Local: South African Perspectives on AI Ethics Risks. <https://ssrn.com/abstract=4240356>.

[97] Musoni, M. Looking into the crystal ball: Artificial intelligence policy and regulation in Africa. <https://ecdpm.org/work/looking-crystal-ball-artificial-intelligence-policy-regulation-africa>.

[98] Adams, R. AI in Africa – Key Concerns and Policy Consideration for the Future of the Continent (2022).

[99] *Ibid.*

[100] UNESCO Recommendation on Ethics of AI, 2022. <https://www.unesco.org/en/articles/recommendation-ethics-artificial-intelligence>.

[101] Olorunju, N. African Algorithmic Governance: Benefit of a Community-based Approach. (2022) <https://researchictafrica.net/2022/04/03/african-algorithmic-governance-benefit-of-a-community-based-approach/>.

[102] Rwanda National AI Policy. <https://www.minict.gov.rw/index.php?eID=dumpFile&t=f&f=67550&token=6195a53203e197efa47592f40ff4aaf24579640e>.

[103] Egypt National AI Strategy. <https://mcit.gov.eg/Upcont/Documents/Publications_672021000_Egypt-National-AI-Strategy-English.pdf>.

## 3.11 Conclusion

The KASE framework applies squarely within the context of South Africa. In addition to these enablers, the interoperability of South Africa's Digital Public Infrastructure (DPI) is an important element for its AIS Stack. Foundation AI models are a form of DPI which underpin AI application infrastructure and can potentially exclude many people from utilising AI.[104] If the DPI is not interoperable and carefully regulated, it remains difficult for any exchanges of data to take place which may inhibit or restrict the development and / or use of AI tools and AI solutions. Secondly, AIS must also be linked to being able to create an AI market where products and solutions are in demand. In the case of South Africa, there are untapped AI markets within the country and outside the country where South Africa can sell its AI products. Without an AI market, it becomes difficult for a country to invest in AI R&D and develop AI solutions.

[104] Ghosh, M. The case for AI foundation models as digital public infrastructure. <https://g-mainak.medium.com/the-case-for-ai-foundation-models-as-digital-public-infrastructure-3ea45896b5bf>.

# 4 AI Sovereignty in India – A Response to the KASE Framework

**Divij Joshi, Lawyer, Doctoral Researcher at University College London**

## Abstract

Artificial Intelligence has a keen hold on the collective imaginations of policymakers in the Global South, including in India. As this set of technologies becomes increasingly influential, questions have been raised about the capabilities and paths towards appropriately developing, using and governing AI systems. One approach, as put forward by Luca Belli, has been to examine AI development and governance from the perspective of ‘sovereignty’, and look at what levers and policies countries can adopt to secure their sovereign interests in AI development. This short response to Belli examines Indian AI policy and governance from the lens of Belli’s ‘Key Enablers of AI Sovereignty’. Further, it interrogates the potential and limitations of sovereignty-based discourses and frameworks, and examines how it might include questions of injustice, equity and democratic participation. Luca Belli’s framework of the Key Enabler’s of AI Sovereignty provides a pragmatic and useful lens through which government institutions and policymakers can understand and respond to concerns about harnessing the capabilities of the unwieldy set of technologies that comprise today’s ‘Artificial Intelligence’ landscape. In examining the Brazilian context, the framework also shows how longer trajectories of media and information governance – telecommunication policy, data protection law, security and adequate infrastructure – each contribute in their own ways to ensuring ‘sovereignty’, which, as per the terms of the paper, is understood as the “capacity of a given country to understand, develop and regulate AI systems.” In this short response, I examine what discourses around AI sovereignty have looked like in the Indian context, where current understandings of ‘sovereignty’ fall short, and what policy considerations might guide AI development in the Global South more broadly.

## 4.1 The 'KASE' in the Indian Context

In defining the component parts of the Key Enabler's of AI Sovereignty (or 'KASE'), the paper mentions data governance and algorithmic governance, strong computational capacity, meaningful connectivity, reliable electrical power, a digitally literate population, solid cybersecurity, and an appropriate regulatory framework. These are interlinked components conceptualised as enabling sovereignty across an 'AI Sovereignty Stack', highlighting that supply chains and infrastructures for AI are essential for establishing meaningful sovereignty. Importantly, this 'stack' is not merely a bundling of technical components, but considers governance and regulation as key components.

In this section, I examine how law and policy on AI in India has developed with a view to each of the KASE, and moreover, whether these are necessary and sufficient indicators of sovereignty in the Indian context.

AI has emerged as a policy concern for the central and state governments in India somewhat sporadically over the last five years, with overlapping policy-building initiatives led by a Government of India research and planning unit, Niti Aayog, as well as by the Ministry of Electronics and Information Technology,[105] and state-level interventions like those by governments in Telangana,[106] Tamil Nadu[107] and Karnataka.[108]

In many of these policy documents, like the National Strategy on Responsible AI,[109] or the Department of Telecom's report on building an Indian AI Stack,[110] **considerations of infrastructure** are paramount – they

[105] Ministry of Electronics and Information Technology, Government of India, 'Artificial Intelligence Committee Reports', (2020), <https://www.meity.gov.in/artificial-intelligence-committees-reports>.

[106] Government of Telangana, 'Telangana AI Framework', <https://startup.telangana.gov.in/wp-content/uploads/2021/04/AI-framework.pdf>.

[107] Information Technology Department, Government of Tamil Nadu, 'Tamil Nadu Safe and Ethical AI Policy', (2020), <https://elcot.in/sites/default/files/AIPolicy2020.pdf>.

[108] For a list of initiatives taken by the Government of Karnataka, see <https://indiaai.gov.in/ministries/government-of-karnataka?initiative=centre-of-excellence-in-data-sciences-and-artificial-intelligence>.

[109] 'National Strategy on Artificial Intelligence, NITI Aayog' <https://niti.gov.in/national-strategy-artificial-intelligence>.

[110] 'Indian Artificial Intelligence Stack', Department of Telecommunications, (2021) <https://www.tec.gov.in/pdf/Whatsnew/ARTIFICIAL%20INTELLIGENCE%20- %20INDIAN%20STACK.pdf>.

stress the need for building computational resources and infrastructure such as data centres, some of which may be realised through initiatives like the 'National Cloud' system being developed by the National Informatics Centre,[111] or through the incentives provided[112] to cloud service platforms to develop data centres in India, which have also been incorporated into policies like the Draft National Data Centre Policy[113] and (draft) E-Commerce Policy.[114]

Domestic semiconductor chip manufacturing – central to any computing hardware industry – has only recently become a focus of industrial policy, two decades after the domestic industry collapsed. Connectivity, another part of the KASE, which could enable access to AI-based online services, has improved significantly over the last decade,[115] particularly in terms of mobile internet coverage. This increase dovetails with strong network neutrality rules adopted in 2017, which have prevented certain forms of anti-competitive behaviour from telecom networks, while still allowing for high growth and investment, including in rural areas.[116]

However, infrastructural concerns are still prevalent across these areas, from the stability of electricity to the computational infrastructure required for training AI models.[117] There are clear disparities in the ability to access infrastructure required to develop or use AI applications, and increasing concentration of internet,

[111] National Informatics Centre, 'National Cloud', <https://www.nic.in/servicecontents/national-cloud/>.

[112] For example, in the Draft E-Commerce Policy, 2019, the Government of India claimed that data localisation would provide a boost to the development of data centres. Various state governments like Karnataka and Maharasthra have provided tax incentives, eased land purchase and provided subsidies for electricity for data centre development. *See* for eg., Government of Karnataka, 'Karnataka Data Centre Policy, 2022', <https://itbtst.karnataka.gov.in/storage/pdf-files/Data%20Center%20Policy.pdf>.

[113] Ministry of Electronics and Information Technology, Government of India, 'Draft Data Centre Policy', (2020) <https://www.meity.gov.in/writereaddata/files/Draft%20Data%20Centre%20Policy%20-%2003112020_v5.5.pdf>.

[114] Ministry of Commerce, Government of India, 'Electronic Commerce in India: Draft National Policy Framework', <https://www.medianama.com/wp-content/uploads/Draft-National-E-commerce-Policy.pdf>.

[115] According to the Telecom Regulatory Authority of India, India has approx. 846 million broadband internet subscribers, of which 813 million are wireless or mobile internet subscribers. See TRAI Telecom Subsription Data, Q1 2023 <https://www.trai.gov.in/sites/default/files/PR_No.46of2023_0.pdf>.

[116] Telecom Regulatory Authority of India, 'Recommendations on Network Neutrality, (28 November, 2017), <https://www.trai.gov.in/sites/default/files/Recommendations_NN_2017_11_28.pdf>.

[117] Husanjot Chahal et. al., 'Mapping India's AI Potential', Centre for Security and Emerging Technology (2021).

computational and data infrastructure in a small number of domestic and international corporations.

In particular, there is a concerning amount of market concentration and lack of public alternatives in various areas of infrastructure development – including a concentration in mobile internet provision, as well as the dominance of private cloud providers in data storage and cloud services, which can be detrimental to goals of strategic autonomy.[118] For example, according to TRAI estimates, the top five service providers constituted 98.37% market share of the total broadband subscribers at the end of March 2023. Of these, Reliance Jio Infocomm has by far the largest market share, around 439 million subscribers.[119]

**Access to data** to train machine learning models, in particular, has been a predominant concern for AI policy in India. A number of policy documents posit data as the 'oil' or raw resource for developing AI in India.[120] However, they also note the lack of useable 'raw data' from citizens, as well as the potential to collect such information for use in the AI supply chain.

The Government of India is taking steps to address this apparent gap in the availability of 'raw data' for AI, both through policy mechanisms around data governance, like the Non-Personal Data Policy proposals,[121] as well as through infrastructural interventions like those made in the National Health Stack[122] and financial Account Aggregator systems,[123] which intend to 'open up' various forms of data for reusage in AI supply chains.

These proposals present some options on how datasets for AI development can be created outside the stronghold of 'big tech'

[118] *Id;* Smriti Parsheera and Vishal Trehan, 'A Structural Analysis of the Mobile Telecommunications Market: Exploring the Jio Effect' <https://publications.clpr.org.in/the-philosophy-and-law-of-information-regulation-in-india/chapter/a-structural-analysis-of-the-mobile-telecommunications-market-exploring-the-jio-effect/>.

[119] *See* note 11.

[120] Cf. Draft E-Commerce Policy, 2019; Non-Personal Data Policy, 2020.

[121] Ministry of Electronics and Information Technology, 'Report by the Committee of Experts on Non-Personal Data Governance Framework', (2020), <https://static.mygov.in/rest/s3fs- public/mygov_160922880751553221.pdf>.

[122] Niti Aayog, 'National Health Stack – Strategy and Approach', (2018) <https://abdm.gov.in:8081/uploads/NHS_Strategy_and_Approach_1_89e2dd8f87.pdf>.

[123] India Stack, <https://indiastack.org/data.html>.

corporations, and indeed present alternative norms for community-centred data governance. However, they fail to address that access to data for AI development must be nuanced and tailored to specific use-cases (purpose limited) and minimally intrusive of privacy,[124] instead of the maximalist approaches taken in data access and sharing proposals.

For example, the policy imperative to make data available for machine learning projects does not take into account mechanisms to audit the appropriateness of datasets, concerns around bias, diversity and representation in the data used, or substantiating claims about how maximising 'data sharing' can promote fairer and more localised AI development.

These proposals replicate issues at the heart of inequities in the development of AI, namely, considering data about people and communities as 'raw resources' for the development of technologies (such as Large Language Models or Facial Recognition Systems which scrape web data), without allowing individuals or communities to have a say in how their data traces are collected or used. Indeed, activities like web-scraping and text and data mining of personal information appear to be condoned by policymakers in India, as apparent from exemptions for using 'publicly available' online data in the Digital Personal Data Protection Act, 2023, enacted in August.[125]

**Governance and ethics concerns** around AI are also emphasised in high-level policy documents. In June 2023, the Telecom Regulatory Authority of India outlined a proposal for an independent AI Regulator,[126] charged with ensuring the 'responsible' development of AI, in line with ethical concerns of fairness, transparency and accountability.

The National Strategy on AI, as well as the policy on Non-Personal Data Governance (which is concerned with the availability and use

[124] 'Fairer Machine Learning in the Real World: Mitigating Discrimination without Collecting Sensitive Data – Michael Veale, Reuben Binns, 2017' <https://journals-sagepub-com.libproxy.ucl.ac.uk/doi/full/10.1177/2053951717743530>.

[125] Section 3, Digital Personal Data Protection Act, 2023, <https://www.meity.gov.in/writereaddata/files/Digital%20Personal%20Data%20Protection%20Act%202023.pdf>.

[126] Telecom Regulatory Authority of India, 'Recommendations on Leveraging Artificial Intelligence and Big Data in Telecommunication Sector', (July, 2023) <https://www.trai.gov.in/sites/default/files/Recommendation_20072023_0.pdf>.

of data for AI applications in India), speak of concerns around bias, transparency and privacy, as well more structural concerns around the availability of datasets for AI.

A number of different approaches have been proposed in this regard, while few have been implemented. Many of these frameworks, while recognising risks such as discrimination and a lack of transparency in AI systems, propose that such risks be dealt through non-binding and voluntary or self-regulatory mechanisms, such as establishing AI ethics principles.

Even as the fundamental right to informational privacy was recognised by the Indian Supreme Court in 2017,[127] a workable data protection law was only enacted in August, 2023, which has several operational and substantive concerns in ensuring privacy in the use of data in the face of AI systems.[128]

Other rights-based concerns around algorithmic discrimination, procedural justice and systematic transparency rights have not been established through any regulatory framework, apart from in some narrow and limited prescriptions on social media algorithms (as in the Information Technology Rules, 2021),[129] or government procurement of algorithmic systems (as in the Tamil Nadu AI Policy, 2020).[130]

## 4.2 Reorienting AI Sovereignty for the Global South

Measuring sovereignty can be a tricky business. The KASE framework speaks to contemporary discourses around 'digital sovereignty', from the lens of state-led interventions in establishing and developing 'AI' as a socially useful technology. Moreover, it attempts to provide not only an indication of how countries are building autonomous capabilities, but equally charts a way for doing so with relative autonomy given the state of global digital supply chains and concerns around 'digital colonialism'.

[127] Justice KS Puttaswamy v Union of India, (2017) 10 SCC 1, (Supreme Court of India).

[128] The Digital Personal Data Protection Act, 2023, (Act 22 of 2023).

[129] The Information Technology (Intermediary Guidelines and Digital Media Ethics Code). Rules, 2021.

[130] Information Technology Department, Government of Tamil Nadu, 'Tamil Nadu Safe and Ethical AI Policy', (2020), <https://elcot.in/sites/default/files/AIPolicy2020.pdf>.

Given the geographical disparities between information-based industries and economies, and the particular histories of global trade and information supply-chains that the AI industry is embedded in,[131] securing sovereign capabilities over AI can be a useful framework through which to mobilise an AI policy agenda, and, at least on paper, the Government of India appears to be cognizant of operationalising AI sovereignty towards these goals, although it remains to be seen how it might be realised.

My broader concern, however, is that while protecting the autonomy of a community (whether defined by ties to nationality or otherwise) to develop an agenda for developing and governing AI is a necessary component of AI policy, current discourses on sovereignty may not be sufficient to safeguard the values that AI currently threatens to harm.[132]

Discourses around digital sovereignty in India, Brazil, and elsewhere in the Global South predominantly emphasise that 'domestic' AI capabilities must be developed autonomously from 'foreign interference' – which is largely seen as national security interests within a cold-war geopolitical framework, or protecting domestic economic interests in the context of the globalised information economy. Policy discourses on AI sovereignty, and digital sovereignty more broadly, have, however, paid little attention to questions of democracy, trust and participation in the development of AI.

Massive data science projects like Large Language Models, biometric recognition models and similar AI systems are overwhelmingly guided by private power and capital, engendering inequitable relations among populations who are being 'datafied' and surveilled,[133] outsourcing the work of labelling and moderation of the data used in AI production to low-wage labour,[134] used to hold sway over the livelihoods of populations affected by automation and

[131] Angelina Fisher and Thomas Streinz, 'Confronting Data Inequality' (2021) 60 Columbia Journal of Transnational Law 829.

[132] *See generally,* 'Algorithmic Accountability: A Primer', Data and Society, (2019), <<https://datasociety.net/library/algorithmic-accountability-a-primer/>; 'AI Risk Management Framework 1.0', NIST, (2023), <<https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf>;.

[133] Lina Dencik and others, 'Exploring Data Justice: Conceptions, Applications and Directions' (2019) 22 Information, Communication & Society 873.

[134] 'Aapti Institute, Just and Equitable Data Labelling towards a Responsible AI Supply Chain', <https://aapti.in/blog/just-and-equitable-data-labelling/> ;.'Exclusive: The $2 Per Hour Workers Who Made ChatGPT Safer' (*Time*, 18 January 2023) <https://time.com/6247678/openai-chatgpt-kenya-workers/>.

shifts in technology-mediated skills,[135] and making consequential decisions about people through the rarified lenses of data science, big data and AI.[136]

Addressing the needs of populations in the global south in the development of AI technologies requires addressing multiple layers of 'AI' – a model for a stack, if you will, which addresses not only the technological and regulatory aspects, but also the multiple scales through which AI development is structured. This requires attention not only to fortifying domestic information industries to keep them globally and domestically competitive, or to protect strategic state interests, but to understand whether and how AI might be responsive to the needs of communities involved in its production and use.

In the context of technology development enmeshed in global capital flows and geopolitical agendas, it also requires attention to strategies for cooperation and collaboration across borders, including mechanisms for reducing concentration of power in big tech, and ensuring that AI development and deployment is not used as a strategy for geopolitical domination.[137]

How might we develop an AI stack that privileges fair labour practices in data labelling and content moderation? How can we reorient sovereignty towards reclaiming decision-making power away from Big Tech and climate polluting data industries and towards addressing the real, contextual needs of people in various contexts? Much of AI policy in India, and the discourses around digital sovereignty globally, fail to address this.

## 4.3 Conclusions

Addressing the inequities which are entwined with the development of contemporary AI systems should feature in the terms of any discussion on 'sovereignty'. Some established frameworks have

[135] Matthew Cole, '(Infra)Structural Discontinuity: Capital, Labour, and Technological Change' (2023) 55 Antipode 348.

[136] Mireille Hildebrandt, 'Privacy as Protection of the Incomputable Self: From Agnostic to Agonistic Machine Learning' (2019) 20 Theoretical Inquiries in Law 83.

[137] *See generally,* Paula Ricaurte, 'Data epistemologies, the coloniality of power, and resistance', 20 *Television & New Media* 4 (2019) 350; Angelina Fischer and Thomas Streinz, 'Confronting data inequality', *Colum. J. Transnat'l L.* 60 (2021) 829.

already demonstrated a path towards more holistic and community-centred governance of data – and data science based projects.

These include, most prominently, frameworks around indigenous data sovereignty,[138] which provide the conceptual tools for decolonising AI design and data governance frameworks by recognising power imbalances inherent in both state-led and private projects for developing AI infrastructure for and by indigenous communities around the world.[139]

Apart from their specific application to indigenous peoples, these frameworks also indicate dimensions of 'sovereignty' that emphasise sensitivity to local context, and responsibility to communities who are impacted by their use. If we are to pursue meaningful sovereignty to understand, develop and regulate AI systems, such frameworks provide us with the vocabulary to demand that we put the interests of workers, impacted communities and users foremost.

[138] Maggie Walter and others, 'Indigenous Data Sovereignty in the Era of Big Data and Open Data' Australian Journal of Social Issues 2021; 56: 143–156.

[139] Lewis, Jason Edward, ed. 'Indigenous Protocol and Artificial Intelligence Position Paper', (2020) Honolulu, Hawai'i: The Initiative for Indigenous Futures and the Canadian Institute for Advanced Research (CIFAR).

# PART 2

## What do AI Transparency and AI Accountability mean?

# 5 Artificial Intelligence: A Theoretical Analysis of Regulatory Models

**Rolf Weber, University of Zurich, Faculty of Law**

## Abstract

Artificial intelligence offers many benefits but also causes some risks. So far, transparency and accountability have often been seen as appropriate "countermeasures" against negative impacts. But a theoretical analysis of these concepts shows that new regulatory models, for example auditability and observability, are better able to avoid undesirable algorithmic data processing and unjust power imbalances. Thereby, soft law instruments containing normative guidelines should complement governmental regulations.

***Keywords:*** *Accountability, auditability, observability, soft law, transparency*

## 5.1 Transparency

### 5.1.1 Notion

Artificial intelligence (AI) offers many benefits but also causes some risks; therefore, the question arises how mitigation measures should be designed. In the past, transparency has often been judged as appropriate remedy. Indeed, already more than 100 years ago (in 1913), Brandeis wanted to make visible the opaque and hidden information, with the objective of creating truth that could enable control and serve as a "disinfectant".[140]

Transparency is usually assessed as encompassing characteristics such as clarity, accuracy, accessibility and truthfulness. These elements are important in the AI context. As in other societal segments, transparency can enable access to the information necessary for the evaluation of opportunities and costs of operations and exchanges. Such an understanding of transparency links information disclosure to visibility, insight, and effective regulatory judgement.[141] This essay

[140] Louis Brandeis, The Other People's Money and How the Bankers Use It, New York 1914, 92.

[141] This essay is based on: Rolf H. Weber, Transparency on Digital Platforms, Weblaw Jusletter, August 31, 2023; this publication analyzing a comparable digital appearance will not be cited anymore further on.

questions the assumption that transparency is sufficient to combat AI challenges and proposes to apply additional regulatory models.

Often transparency is differentiated into three main pillars, namely (i) procedural transparency, (ii) decision-making transparency and (iii) substantive transparency:[142]

*(i) Procedural transparency* encompasses rules and procedures in the operation of legal entities that must be clearly stated, have an unambiguous character and are publicly disclosed. The rules should also make the process of governance and law-making accessible and comprehensible for the public.

*(ii) Decision-making transparency* can be seen as reasoned explanations for decisions that, together with public scrutiny, are able to strengthen the institutional credibility and legitimacy of decisions.

*(iii) Substantive transparency* is directed at the establishment of rules containing the desired substance of revelations, standards and provisions which avoid arbitrary or discriminatory decisions; substantive rules often include requirements of rationality and fairness.

In the AI context, all three elements are relevant. The concerned persons need to know how the data processing is conducted (procedure), who is taking decisions and what material standards are applied. The compliance with the three elements also impacts the assessment of the below discussed accountability.

## 5.2 Challenges

But the concept of transparency having become an essential regulatory element mainly in financial markets and consumer laws, is increasingly exposed to challenges and critical analyzes. Echoing these voices, transparency is partly seen as policy panacea.[143] AI data processing is exposed to false binaries between secrecy and openness, to strategic occlusions and to market-dominant behavior of big enterprises; these factors influencing the algorithmic matching results can lead to power imbalances.

[142] See Rolf H. Weber, Shaping Internet Governance: Regulatory Challenges, Zurich 2009, 121.

[143] For further details see Bernhard Rieder/Jeanette Hofmann, Towards platform observability, Internet Policy Review 9 (2020), 1, 3–6, <https://doi.org/10.14763/2020.4.1535>.

## 5.3 Comprehensibility

Over the last ten years, regulations in financial markets and consumer segments have substantially increased the scope of information duties to be observed by providers of goods and services.[144] Examples are the extensive information requirements for capital markets prospectuses and the specific (hardly understandable) descriptions for medical products. As mentioned, Brandeis attributed the characteristics of "sunlight" and "disinfectant" to the transparency principle; however, since the detailed disclosure often goes too far, the recipient does not anymore understand its key message.

Transparency should address the way how information is delivered in order to optimize the outcome of the informational process. The basic objectives of transparency require robust and general rules; this principle is now enshrined in article 12 para. 1 GDPR; information must be given "in a concise, transparent, intangible and easily accessible form, using clear and plain language".[145] If the information is clear and straightforward, the addressee will be able to fully understand it (so-called comprehensibility).[146] The GDPR requirement is very appropriate, however, the reality in the AI environment shows that the chosen information approach often does not comply with the GDPR requirements (Article 12).

## 5.4 Mandated Disclosure

In the business-oriented context, the increasing number of information obligations has been mainly criticized by representatives of the law and economics discipline under the heading of "mandated disclosure paradigm".[147] Apart from the hidden costs caused by such kind of disclosure (for example detailed information obligations in the Artificial Intelligence Act of the EU[148]), Ben-Shahar & Schneider argue

[144] Rolf H. Weber, The Disclosure Dream – Towards a New Transparency Concept in Consumer Law, EuCML 2023, 67–68.

[145] General Data Protection Regulation (GDPR) 2016/679 of April 2016, OJ 2016 L 119 of 4 May 2016.

[146] See also Rolf H. Weber, From Disclosure to Transparency in Consumer Law, in: K. Mathis/A. Tor (eds.), Consumer Law and Economics, Cham 2021, 73, 79–81.

[147] For a general assessment see Omri Ben-Shahar/Carl E. Schneider, More Than You Wanted to Know: The Failure of Mandated Disclosure, Princeton 2014.

[148] Proposal for a Regulation laying down harmonised rules on artificial intelligence (Artificial Intelligence Act, AIA) of April 21, 2021, COM (2021) 206 final, <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206>. The final adoption of the AIA is scheduled for fall 2023.

that the mandated disclosure would exacerbate inequality, impair consumers' decisions and deter lawmakers from adopting better regulations;[149] and these authors add that the provided information whether individually aggregated or based on advice "will not adequately help the naïves in their dealings with the sophisticated".[150]

Even if these statements could be partly contested, it appears to be doubtful that the information addressees indeed read and understand the mandatorily provided information. Distributed ledger technologies aggravate the problem: platform users are often not able to understand the "IT codes" meaning that for example the disclosure of mathematical formulas constituting a smart contract do not lead to an informed addressee. Therefore, a potential way to assist individuals in making better decisions would rather be to direct choices through smart incentives without mandating a certain outcome.

## 5.5 Information Overload

The transparency principle is also confronted with the issue of information overload. Looking from a societal perspective, too detailed information requirements could have two negative effects:[151]

- The sheer volume and intensity of information leads to a confusion effect since the recipients are not anymore able to cope with all information details and loose the necessary overview in respect of the disclosed data.
- The permanent delivery of (similar) information causes a Kassandra effect; even if the recipients take note of the information, its contents is no longer seen as being serious and reliable.

The general wisdom that overconsumption of information can have negative effects or even be risky also applies in respect of detailed disclosure requirements:[152] (i) Over-information consumes working and leisure time on both sides of an informational relationship.

[149] See Ben-Shahar/Schneider (supra note 8); for a more detailed discussion see Weber (supra note 7), 75 and 77–78.

[150] Omri Ben-Shahar/Carl E. Schneider, The Failure of Mandated Disclosure, University of Pennsylvania Law Review 159 (2011), 647, 748.

[151] See Weber (supra note 7), 79–80 with further references.

[152] See Niklas Luhmann, Die Gesellschaft der Gesellschaft, Frankfurt 1997, 1090, 1097 and 1102.

(ii) Attention is a scarce resource; a person cannot dispose of this resource in an unlimited way. (iii) Over-information increases the risk that messages or data being spread out are considered to be redundant.

Notwithstanding the fact that detailed information provisions can constitute a certain value for persons having a broader expertise (academics, lawyers), it should not be underestimated that the balancing of interests remains difficult in relation to addressees not experienced in the AI services context. Incomplete disclosure leaves people ignorant, but complete disclosure creates overload problems;[153] as a consequence, the regulator should recognize that "less is more" even if it cannot be excluded that "less is not enough".[154]

## 5.6 Accountability

### 5.6.1 Notion

Accountability, often being called a "twin" of transparency, encompasses the obligation of one person or legal entity to give account of, explain and justify the undertaken actions or decisions to another person in an appropriate way.[155] Accountability is a pervasive concept, including political, legal, philosophical, and other aspects, each of them casting a different shade on the meaning of the terms. Checks and balances as emanation of accountability constitute a prerequisite for legitimacy and a key element of any governance discussion.

As a fundamental principle, accountability concerns itself with power and power implies responsibility. Therefore, accountability can be framed among three elements,[156] namely (i) the provision of information in a timely manner, (ii) the introduction of standards that hold governing bodies accountable, and (iii) the implementation of mechanisms of sanction. In addition, accountability needs to include the democracy element of global governance if the outcome

[153] Weber (supra note 7), 79–80.

[154] Ben-Shahar/Schneider (supra note 11), 647.

[155] Weber (supra note 3), 133 with further references.

[156] See Rolf H. Weber, Internet Governance at the Point of No Return, Zurich 2021, 70.

of the decision-making processes should be acceptable to civil society in general.[157]

## 5.7 Challenges

Regulatory approaches seeking to create accountability in the AI context have to tackle the relevant issues by "opening the black box" of algorithmic decision-making.[158] However, accountability should extend beyond oversight of algorithms and behavioral conduct. In addition, behavior cannot be reduced to conduct since tight integration of data collections and targeted "intervention" in form of "surveillance capitalism"[159] has produced a market form that is unimageable outside the digital milieu. As a consequence, the risingpower of AI providers makes it necessary to assess what kind of accountability must be applied in order to understand the processes and their consequences in more detail.[160]

A specific challenge to accountability is the occurrence of opacity[161] being an obvious concern that may stem from the increasing "mismatch between mathematical optimization in high-dimensionality characteristics of machine learning and the demands of human-scale reasoning".[162] The applicable techniques usually develop decision models inductively and learn programs from data.

Since many variables come into play, academics argue that the developed algorithms are not easily" legible" in daily life.[163] Consequently, transparency in the sense of reconstructing the procedure of algorithmic decision-making often does not lead to an informative outcome. Even if regulators were given access to data centers and source code, the above discussed comprehensibility

[157] Rieder/Hofmann (supra note 4), 6.

[158] The black box problem is fundamentally described by Frank Pasquale, The Black Box Society. The Secret Algorithms That Control Money and Information, Cambridge MA 2015.

[159] Term introduced by Shoshanna Zuboff, The Age of Surveillance Capitalism. The Fight for a Human Future at the New Frontier of Power, New York 2019, 15.

[160] To the autonomy and power elements in the accountability context see also Weber (supra note 17), 71.

[161] Rieder/Hofmann (supra note 4), 6–7.

[162] Jenna Burrell, How the machine 'thinks': Understanding opacity in machine learning algorithms, Big Data & Society 3 (2016), 1–2.

[163] Ansgar Koene et al., A governance framework for algorithmic accountability and transparency, European Parliamentary Research Service Study, April 2019, 31–32.

would not be straightforward in view of complex code designs and involved machine learning. In addition, the existence of different programming languages and execution environments adds further complications.[164]

## 5.8 Alternative Approaches and Regulatory Models

### 5.8.1 Introduction

Transparency has a long tradition as a regulatory model. The assumption, however, that transparency is able to reveal the truth by reflecting the internal reality of an organization is not fully reflected in reality. As mentioned, research on transparency has shown that this principle does more and different things than shedding light on what is hidden. The visibility of an entity offering AI services and its procedures is not simply a disclosure of pre-existing facts, but a process that implies its own perspective.

Therefore, transparency as well as accountability should not be regarded as a state or a “theme” but as the practice of deciding what to make present (i.e. public and transparent) and what to keep confidential.[165] Creating visibility and insights is a specific process which involves choices about what specifically should be exposed and how, what is relevant and what can be neglected, which elements should be shown to whom and how the visible aspects could be interpreted.[166] Potential elements being able to design such a process are auditability and observability.

Apart from the search of suitable regulatory approaches the appropriate normative models need to be analyzed. Insofar soft law instruments developed by the concerned stakeholders of civil society merit special attention.

## 5.9 Auditability

An improvement of transparency and accountability can be achieved by extended auditability requirements if the respective provisions

[164] Paul Dourish, Algorithms and their others: Algorithmic cultures in context, Big Data & Society 3 (2016), 1, 4.

[165] Rieder/Hofmann (supra note 4), 5.

[166] Weber (supra note 5), 70.

overcome an insufficient understanding of algorithms and platform architectures.[167] In order to reach the (theoretical) transparency objective, it would be necessary to develop an institutionalized mechanism for the verification of AI-provided information or data. The respective efforts are done under the heading of "auditability".

Several aspects need to be considered in the implementation of auditability principles:[168] (i) The creation of an intermediary (public or private sector entity) that audits data provided by large online platforms can ensure the accuracy of data. (ii) By bundling the auditing process through centralized auditing intermediaries, the exposure of sensitive private data to as few actors as possible is limited. (iii) By distancing the audit process from the regulator that is asking for data ensures that regulatory action does not overstep its bounds. (iv) By limiting the number of points through which the online platforms need to interact with outside intermediaries limits potential security risks that could arise from providing access to a wide variety of systems. (v) Having numerous regulators involved in auditing is likely to create unnecessary and redundant processes. (vi) Organizing auditing of transparency data through an external auditing intermediary ensures that even regulators without the capacity to organize audits themselves still may have access to such a system through auditing intermediaries.

The most important question about an auditing intermediary concerns the decision of whether such an intermediary would be public, private or somewhere in between.[169] Such an institution could be created within the context of the recently adopted EU Digital Services Act (DSA).[170] A further challenge raised by the proposal of auditing intermediaries is how much access to data these intermediaries would actually need. In particular, it must be avoided that auditing intermediaries are misused by authoritarian countries for strategic national interests.[171]

[167] See also Mike Annany/Kate Crawford, Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability, New Media & Society 20 (2018), 973–974.

[168] For further details see Ben Wagner/Lubos Kuklis, Establishing Auditing Intermediaries to Verify Platform Data, in: M. Moore/D. Tambini (eds.), Regulating Big Tech, Oxford 2021, 169, 172–173.

[169] Wagner/Kuklis (supra note 29), 174.

[170] Regulation 2022/2065 of 19 October 2022 on a Single Market for Digital Services, OJ 277 L 1 of 27 October 2022.

[171] Wagner/Kuklis (supra note 29), 174–175.

## 5.10 Observability

A further approach proposes to realize a concept of observability as a pragmatic way of thinking about the means and strategies necessary to hold AI providers accountable.[172] Unlike transparency being normally described as a state that may exist or not, observability emphases the conditions for the practice of observing in a given domain. While observability incorporates similar regulatory goals as transparency, it also partly deviates, most importantly by understanding accountability as a complex, dynamic "social relation".[173] Observability should be a mechanism that can overcome the lack of sensitivity for fundamental power imbalances, strategic occlusions, and false binaries between secrecy and openness.

The challenges raised need to be addressed in a broad way, beginning with the question of how large-scale, transnational environments that heavily rely on technology as a mode of governance can be assessed.[174] The concept of observability seeks to develop concrete actions in respect of (i) how people need to be treated in the digitized environment, (ii) how connections between participants are made and structured, and (iii) which outcomes should be achievable.[175]

In the academic literature, the concept of observability starts with the recognition of a growing information asymmetry between AI services providers and civil society. The frequently given data monopoly situation deprives society of a crucial resource for producing knowledge about itself. The deep political and social repercussions reflect the need to implement broader forms of social accountability.[176] The concept of observability should be based on public interest as a normative horizon for assessing and regulating the societal challenges. In the context of the public sphere, public interest encompasses the protection of human rights such as the

[172] A broad discussion of the observability concept is offered by Rieder/Hofmann (supra note 4), 9–18.

[173] See Mark Bovens, Analysing and Assessing Accountability: A Conceptual Framework, European Law Journal 13 (2007), 447, 450.

[174] Rieder/Hofmann (supra note 4), 9–10.

[175] See also José von Dijck/Thomas Poell/Martin de Waal, The Values in a Connective World, Oxford 2018, 158.

[176] Rieder/Hofmann (supra note 4), 9–10.

freedom of expression and the freedom of information, fostering cultural and political diversity throughout the whole society.[177]

Furthermore, the principle of observability reflects the acknowledgment that the volatility of AI solutions requires continuous observation. If terms of services contracts would be made available as machine-readable documents, the ongoing observation and interpretation of AI-related activities could be facilitated.[178] Another factor concerns the availability of interfaces that provide continuous access to relevant data. Thereby, questions of how data and analytical capacities are made available, to whom, and for what purpose need to be tackled.[179]

Observability requires a critical audience. But the capacity for critique must be broader than "only" a critical attitude. Moreover, frameworks for data access should be linked to a cultivation of a robust civil society. Therefore, observability as a social relation makes scrutiny of realized transparency by a specific forum necessary.[180]

Regulating AI providers with the objective of increasing observability does mean working towards structured information interfaces between them and society. Such kind of regulation requires engaging with the specific properties of algorithmic systems and the co-produced nature of AI results. The complex interactions between technical design, terms of service, and often large numbers of both users and "items/issues" have the consequence that the existing processes are conceptually insufficient.[181] AI providers should become subject to public interest requirements as a normative benchmark; elements could consist of risk control measures, auditability reviews, behavioral rules and strict responsibility obligations.

## 5.11 Soft Law Instruments

As far as the normative rule-making models are concerned, a fresh thinking appears to be necessary. Existing and future governmental

[177] See also José van Djick, Governing digital societies: Private platforms, public values, Computer Law & Security Review 36 (2020), 1, 3.

[178] Rieder/Hofmann (supra note 4), 13–14.

[179] Von Djick (supra note 38), 3.

[180] Bovens (supra note 34), 450.

[181] See also Philip M. Napoli, Social media and the public interest: Governance of new platforms in the realm of individual and algorithmic gatekeepers, Telecommunications Policy 39 (2015), 751 et seq.

regulations should be complemented by self-regulatory and co-regulatory mechanisms that specify the general legal framework in more detail. Soft law is playing an increasingly important role in the digitized world,[182] since it has the advantage of being usually developed by the concerned community members (for example market participants, consumer organizations) and of having a cross-border reach without restrictions of national boundaries.[183]

As an example, the UN IGF Coalition on Platform Responsibility has presented a "Model Framework for Meaningful and Interoperable Transparency for Digital Platforms" at the occasion of the Internet Governance Forum in December 2022;[184] the "Model Framework" refers, as the name says, to digital platforms, but AI services have parallel characteristics in my respects. As a key objective it must be made sure that quantitative data can be assessed from a qualitative perspective; therefore, digital platforms (or AI providers) should make available data sets including qualitative information on (i) which content was reported, (ii) which measures were taken by the platform (or AI provider), (iii) which procedures were adopted (maintenance, removal, depriorization, etc.), (iv) to what extent due process requirements were applied and (v) what the consequence of user appeal has been.[185]

The "Model Framework" proposes standardized and shared rules:[186] From a *substantive* perspective, platforms (or AI providers) should share detailed and intelligible information on (i) their content moderation rules, (ii) the functioning of automated algorithmic moderation systems, and (iii) due process procedures. From a *methodological* perspective, platforms (or AI providers) should (i) collectively standardize the information provision, (ii) make data continuously available in an interoperable, understandable and machine-readable format as audited by third parties, and (iii) publish their initiatives regarding the identification and prevention of biases

[182] For further details see Rolf H. Weber, Sectoral Self-Regulation as a Viable Tool, in: K. Mathis/A. Tor (eds.), Law and Economics of Regulation, Cham 2021, 5, 26–27.

[183] Weber (supra note 43), 27–28 with further references.

[184] See Luca Belli/Yasmin Curzi/Clara Almeida/Natália Couto/Roxana Radu/Rolf H. Weber/Ian Brown, Towards Meaningful and Interoperable Transparency for Digital Platforms, UN IGF 2022, <https://www.intgovforum.org/en/filedepot_download/57/23886>.

[185] Belli et al. (supra note 45), 7.

[186] Belli et al. (supra note 45), 7.

in their algorithms. As mentioned, these principles could comparably be designed in respect of the AI services providers' commitments.

In addition, the implementation of complaints-handling processes is imperative. An independent body of experts must be established being capable of assessing the different potential kinds of complaints raised by the concerned persons.

## 5.12 Outlook

The ongoing discussions about transparency and accountability in the AI environment reveal that at first instance major emphasis should be put on the quality of information and not on the extension of the quantity of information (partly done in national regulations). As shown, only a meaningful understanding of transparency and accountability can serve as an effective check in respect of power structures. Not more information is needed, but a better structured disclosure of data becomes imperative. Salience matters when certain information is essential for the individuals or general welfare.[187]

In view of far-reaching and partly not in detail foreseeable developments of AI applications, a three-dimensional concept of transparency/accountability merits to be implemented: (i) The first dimension refers to institutional aspects, i.e. procedures and decision-making. (ii) The second dimension of transparency constitutes the substantive backbone of the regulations. (iii) The third dimension is accountability of actors for rebuilding confidence in the market system.[188]

Furthermore, an appropriately targeted transparency/accountability should encompass additional regulatory models such as the auditability and the observability concepts. Information contents must be designed in view of the potential addressees and of the used AI-mechanisms, thereby leading to their improved empowerment:[189] (i) Individuals being subject to AI services should be informed about how (personal) information will be used and organized by the AI provider and about decisions related to content or account that may occur. (ii) Civil society or the general public needs information

[187] See also Rieder/Hofmann (supra note 4), 23.

[188] For further details see Weber (supra note 3), 140–143, 147.

[189] See Belli et al. (supra note 45), 6–7, in respect of digital platforms.

about the functioning and the algorithmic instruments applying AI methods. (iii) Regulatory bodies, public supervisors and other auditing bodies are to be informed about the implementation of protection measures and the compliance with existing regulations.

# 6 A conceptual framework for AI supply chain regulation

**Ian Brown, Visiting Professor, Centre for Technology & Society, FGV Law School, Rio de Janeiro**

## Abstract

Building on existing work on the regulation of components of AI supply chains, we develop a conceptual framework for policymakers and regulators to apply different responsibilities in the regulation of AI systems to their constituent parts. This approach complements requirements from a range of existing legal frameworks including data protection, copyright, equality and non-discrimination, and contractual liability. We describe a framework focusing on principles of transparency, incentivisation, efficacy and accountability. To support this framework, regulation will need to require the use of various transparency mechanisms that enable a flow of critical information and modes of redress up and down an AI system's supply chain and identify new ways to incentivise these practices. The advent of general-purpose AI systems (such as OpenAI's GPT-4) likely to be present in many supply chains complicates the challenge of allocating responsibility. We discuss how various aspects of these nascent systems (including who is designing them, how they are released and what information is made available about them) may impact the allocation of responsibilities for addressing potential risks. While jurisdictions including the US and UK are focusing regulation on customer-facing businesses, some firms supplying services incorporating AI components directly to end-users will not have the power, access or capability to address or mitigate all risks or harms that may arise from their supply chain as a whole. Finally, we discuss some of the challenges that open-source technologies raise for AI supply chains. We suggest policymakers focus on how AI systems are released into public use, which can inform the allocation of responsibilities for addressing harms along a supply chain.

**Acknowledgements:** This is an edited and updated report on research undertaken by the author with the UK's Ada Lovelace

Institute, with financial support from UK Research and Innovation. The author would like to thank the Ada Lovelace Institute's Elliot Jones and Andrew Strait for their substantive contributions; Luca Belli and Centre for Technology and Society colleagues for their input; and Reuben Binns, Connor Dunlop, Hamed Haddadi, Natali Helberger, Jat Singh, Chris Marsden and the anonymous reviewer for their helpful comments.

## 6.1 Introduction

Developers and deployers of AI systems have a variety of distinct responsibilities for addressing risks through their lifecycle, from problem definition to data collection / labelling / cleaning, model training and fine-tuning, then testing and deployment of an AI system. These activities are potentially carried out by different companies in a supply chain. To ensure AI systems are safe and fit for purpose, actors in their supply chains must be accountable for evaluating and mitigating these different risks.

Every AI system will have a different supply chain, with variations depending on the sector, the use case, whether the system is developed in-house or procured, and how the system is made available to those who use it (e.g. via an application programming interface (API), or made available via a hosted platform). Actors along each chain will have differing but overlapping obligations to assess and mitigate these risks, and some actors will be more responsible than others. This makes developing a single framework for accountability along supply chains for AI systems challenging.

Previous work has analysed how supply chain components would be regulated under the European Union's proposed AI Act,[190] and more broadly how regulation can be applied to information society services with complex supply chains suffering from a "many hands" problem.[191] Based on a rapid review of academic and grey literature, this article analyses which actors should be primarily responsible

[190] Alex Engler and Andrea Renda, 'Reconciling the AI Value Chain with the EU's Artificial Intelligence Act' (Centre for European Policy Studies 2022), pp. 2–3, <https://www.ceps.eu/ceps-publications/reconciling-the-ai-value-chain-with-the-eus-artificial-intelligence-act/>.

[191] Natali Helberger, Jo Pierson and Thomas Poell, 'Governing Online Platforms: From Contested to Cooperative Responsibility' (2018) 34 The Information Society 1.

for different risks in a more complex, real-life AI supply chains, and what mechanisms may allow downstream actors to reach back up through the supply chain to flag issues that they cannot deal with in isolation. We include examples of national approaches from the EU, US, Japan and Singapore. Relevant literature was identified through keyword searching of online databases of academic literature and through snowball sampling via discussions with experts in AI supply chains and risk management.

In the next section we set out our conceptual framework for considering AI supply chains, based around four principles: transparency, incentivisation, efficacy and accountability. The framework considers the information flows necessary to enable actors to assess and remedy harms; what incentives will be needed to encourage them to do so; which actors will be in a position to identify and mitigate risks; and how contractual chains of liability will and will not enable allocation of responsibility, especially between imbalanced actors (such as small software providers and the largest technology companies providing AI services, such as Google, Amazon and Microsoft).

In section 4 we then consider how "general-purpose" or "foundation" AI models, trained on very large quantities of data and applicable to many different tasks, fit into this analysis. Their cross-functionality makes them less amenable to sector-specific regulation. The data and concomitant large-scale computation requirements for training these models is likely to have implications for industry concentration and the market power of the largest providers, already seen in the global cloud computing market which is likely to underpin the creation and provision of these services.

Finally in section 5 we consider the impact on accountability of different release strategies for AI system components, from tightly-controlled services provided via limited "Application Programming Interfaces" to fully open releases of models and the software and data used to create them. More openness can bring benefits, as it increases the ability of a wider range of organisations and experts to audit models, increases the transparency of how models work and brings a broader range of perspectives to bear. It also enables broader participation in the development of complex models, partially addressing concerns about industry concentration. But more open releases can also reduce the technical ability of AI developers to constrain their systems' use or misuse.

## 6.2 A conceptual framework for AI supply chains

Policymakers and regulators must grapple with questions of where to assign distinct responsibilities for addressing the risks of AI throughout an AI system's supply chain. Below, we provide an initial conceptual framework that regulators can follow to determine where responsibilities might apply, which relies on four principles:

- **Transparency:** what information can each actor in a supply chain provide to enable risks to be identified and addressed.[3]
- **Incentivisation:** who is best incentivised to address these risks, and how can regulators create those incentives while minimising the overall costs of fixing problems.
- **Efficacy:** who is best-positioned to most effectively address the risks that can emerge from an AI system (potentially multiple parties working together).
- **Accountability:** how can the use of legal contracts assign responsibilities, and what are the limitations of this method.

## 6.3 Transparency

To ensure effective regulation, regulators and policymakers will need to incentivise transparency and information flow across the supply chain, so that information about and evaluation of systems and potential risks can travel up and down chains, supporting remediation of identified problems.

Mechanisms needed to ensure this flow of information, including via contractual terms and regulatory requirements on all actors in a supply chain, include:

- **Transparency and accountability mechanisms, including model cards, datasheets, etc**. which provide information on an AI model's architecture and the data they were trained on. These 'have the potential to increase transparency and accountability within the machine learning community, mitigate unwanted societal biases in machine learning models, facilitate greater reproducibility of machine learning results, and help researchers and practitioners to select more appropriate datasets for their chosen tasks'.[192]

- **Certifications, audits, impact assessments, technical standards and similar mechanisms**, which give organisations reliable evidence on the trustworthiness of AI systems.[193] They will allow organisations to evaluate and monitor aspects of components that are important to their regulatory duties and their end-users.
- **Sector-specific information-sharing**, like the UK's Cyber Security Information Sharing Partnership, potentially facilitated by regulators. These kinds of fora could also develop voluntary sectoral codes of conduct, building on those envisaged in the GDPR's Articles 40 and 41, and developing standards for certifications.[194]
- **Data required by insurers and regulators**, for example, in the related area of cybersecurity, one US review found 'a lack of data, a lack of expertise, and an inability to scale rigorous security audits have rendered cyber insurers unable to play a significant deterrent role in reducing cybersecurity incidents or exposure to cyber risks.' The review highlights the approach of the Singaporean government in improving this issue: 'developing a standardized taxonomy for describing cybersecurity incidents, creating a database of cybersecurity incidents and their resulting losses, and benchmarking different models of cyber-related losses to support actuarial pricing.'[195]
- **Mechanisms for reporting and remedying faults**, researchers from Stanford's Human-Centered AI project suggested: 'If downstream users have feedback, such as specific failure cases or systematic biases, they should be able to publicly report these to the developer, akin to filing software bug reports. Conversely, if a model developer updates or deprecates a model, they should notify all downstream

[192] ibid.

[193] Centre for Data Ethics and Innovation, 'The Roadmap to an Effective AI Assurance Ecosystem – Extended Version' (2021) <https://www.gov.uk/government/publications/the-roadmap-to-an-effective-ai-assurance-ecosystem/the-roadmap-to-an-effective-ai-assurance-ecosystem-extended-version> accessed 11 March 2023.

[194] L Edwards and M Veale, 'Slave to the Algorithm? Why a "Right to an Explanation" Is Probably Not the Remedy You Are Looking For' (2017) 16 Duke Law & Technology Review 70–80.

[195] Shauhun Talesh, 'Cyber Insurance and Cybersecurity Policy: An Interconnected History' (*Lawfare*, 4 November 2022) <https://www.lawfareblog.com/cyber-insurance-and-cybersecurity-policy-interconnected-history> accessed 23 March 2023.

users' including deployers or end users whose products and services rely on that model.[196]

More broadly, it may be most efficient for a government body to play a cross-sectoral role for information-sharing and learning.[197] In the Netherlands, for example, an algorithm regulator, situated within the Data Protection Authority, 'will identify cross-sector risks related to algorithms and AI and will share knowledge about them with the other regulators. It will also, in cooperation with already existing regulators, publish and share guidance related to algorithms and AI with market parties, clients and governments.'[198] These bodies can collaborate internationally in venues such as the Organisation for Economic Co-operation and Development (OECD) and Council of Europe. Policymakers will also need to consider the impact of trade secrecy on the willingness (or otherwise) of actors to share information about their systems.

So-called 'explainable' AI (XAI) systems may help with allocation of responsibility, in that '[d]eveloping systems that can explain their "thinking" will let lawyers, policymakers and ethicists create standards that allow us to hold flawed or biased AI accountable under the law.'[199] However, some researchers have noted the limitations of current XAI approaches, which can be brittle and change over time.[200]

Finally, regulators and policymakers must acknowledge the limits of transparency. Simply making information about AI systems, data or risks available does not mean that information will be acted on by relevant parties. Regulation must create proportionate incentives for them to do so.

[196] Percy Liang and others, 'The Time Is Now to Develop Community Norms for the Release of Foundation Models' (*Stanford University Human-Centered Artificial Intelligence*, 17 May 2022) <https://hai.stanford.edu/news/time-now-develop-community-norms-release-foundation-models>.

[197] For a greater discussion on AI monitoring, see: Ada Lovelace Institute (2023), *Approaches to government monitoring of the AI landscape,* [Internal briefing for DCMS].

[198] Martijn Schoonewille and others, 'Introduction New Algorithm Regulator and Implications for Financial Sector' *Lexology* (5 January 2023) <https://www.lexology.com/library/detail.aspx?g=3e71f01b-2cb7-4294-b8f2-68ea2ab67261>. Accessed 20 January 2023.

[199] Mason Kortz and Finale Doshi-Velez, 'Accountability of AI Under the Law: The Role of Explanation' (Berkman Klein Center 2017) <https://cyber.harvard.edu/publications/2017/11/AIExplanation>.

[200] de Bruijn, H., Warnier, M. and Janssen, M. (2022) 'The perils and pitfalls of explainable AI: Strategies for explaining algorithmic decision-making', Government Information Quarterly, 39(2), p. 101666. <https://doi.org/10.1016/j.giq.2021.101666>.

## 6.4 Incentives and value chains

Another principle regulators can use is to ask who is best incentivised to address emerging risks in an AI supply chain, while considering the risk of “diffusion of responsibility” among many actors in complex supply chains leading to an insufficient consideration by any of them.[201]

Current corporate practices often do not align with incentives to produce systems that prioritise societal benefit. In interviews with 27 AI practitioners, scholars found a ’deeply dislocated sense of accountability, where acknowledgement of harms was consistent but nevertheless another person’s job to address, almost always at another location in the broader system of production, outside one’s immediate team’.[202]

Interfaces along a supply chain could be strengthened through the use of legal contracts that specify clear responsibilities and increase communication between non-developers: ‘those playing customer roles in the supply chain might routinize asking suppliers for model cards, if the data it was trained on was properly consented, if crowd workers labelling the data were paid an appropriate wage, etc., which is commonplace in supply chains for physical goods’.[203]

## 6.5 Efficacy up and down the AI supply chain

Regulators and policymakers must also consider which actor in a value chain can most easily identify risks, and which actor is best-placed to take action to mitigate them.[204] European civil society organisations have argued that shifting the obligations entirely to downstream users in a supply chain ‘would make these systems less safe’, as they are likely to lack the capacity, skills and access to the model to make any changes. However, they have also argued that downstream companies deploying the system are best-placed to comply with other requirements of the act like ‘human oversight, but

[201] John M Darley and Bibb Latane, ‘Bystander Intervention in Emergencies: Diffusion of Responsibility’ (1968) 8 Journal of Personality and Social Psychology 377.

[202] David Gray Widder and Dawn Nafus, ‘Dislocated Accountabilities in the AI Supply Chain: Modularity and Developers’ Notions of Responsibility’ [2023] Big Data & Society <http://arxiv.org/abs/2209.09780> accessed 17 January 2023.

[203] ibid.

[204] Engler and Renda (n 1) 24.

also any use case specific quality management process, technical documentation and logging, as well as any additional robustness and accuracy testing.'[205] This is because downstream deployers are closer in proximity to the final context in which the system is operating.

## 6.6 Accountability through contracts

Companies offering products and services to the market that contain or are based on AI components will generally bear the legal liability of doing so. Where courts or regulators fine or order compensation payments against such companies, they will in turn need to examine whether their suppliers should be responsible for some (or all) of these remedies. As researchers have observed: 'Apportioning blame within the supply chain will involve not only technical analysis regarding the sources of various aspects of the AI algorithm, but also the legal agreements among the companies involved, including any associated indemnification agreements.'[206]

At a minimum, those firms will need to use contract law to ensure they have all the data they need about the models and systems they make use of to do so effectively.[207] Japan's government is encouraging this by issuing interpretive guidance on AI contracts.[208] In turn, their suppliers will need to ensure they can do the same with all of the components making up the systems they are offering. Similarly, those contracts will need to provide mechanisms by which firms using AI can notify suppliers and request remediation of problems, all the way up the supply chain.

Debate in EU institutions has also highlighted 'the belief that original AI developers will often be larger entities such as tech giants. These larger entities can be assumed to possess more resources and greater knowledge compared to the (arguably smaller) companies that will

[205] Access Now et al., 'Call for Better Protections of People Affected at the Source of the AI Value Chain' (25 October 2022) <https://futureoflife.org/wp-content/uploads/2022/10/Civil-society-letter-GPAIS-October-2022.pdf> accessed 21 March 2023.

[206] ibid.

[207] Engler and Renda (n 1) 15.

[208] MEIT expert group, 'Governance Guidelines for Implementation of AI Principles Ver. 1.1' (Japan Ministry of Economy, Trade and Industry 2021) 35 <https://www.meti.go.jp/english/press/2022/0128_003.html>.

eventually become the providers, as they will place the high-risk AI systems on the market.'[209]

Upstream suppliers will often be larger / more powerful, and downstream deployers will have (very) limited ability to negotiate custom contracts – as already seen with cloud services. This may leave small and medium-sized enterprises (SMEs) in a weak position to determine important aspects of contracts.[210]

## 6.7 General-purpose AI (GPAI) systems

General-purpose AI (GPAI) systems are worth considering as a separate element of an AI supply chain, because they further complicate the ability for regulators to assign responsibilities, and make it more challenging for sectoral regulators to know where their remit should apply.

GPAI systems, sometimes called 'foundation models', 'are characterised by their training on especially large datasets to perform many tasks, making them particularly well-suited for adaptation to more specific tasks through transfer learning. These models – especially those used for natural language processing, computer vision, speech recognition, simulation, and robotics –have become more foundational in many commercial and academic AI applications.'[211] OpenAI's chief scientist Ilya Sutskever has commented: 'These models are… becoming more and more potent. At some point it will be quite easy, if one wanted, to cause a great deal of harm with those models.'[212]

A single GPAI model can be adapted (or 'fine-tuned') for a wide variety of applications, which means:

1. It becomes harder for upstream providers of a GPAI model to understand how it will be used and to mitigate its risks.
2. A much wider number of sectoral regulators will have to evaluate its use.

[209] Engler and Renda (n 1) 23.

[210] J Cobbe and J Singh, 'Artificial Intelligence as a Service: Legal Responsibilities, Liabilities, and Policy Challenges' (2021) 42 Computer Law & Security Review 105573, 43.

[211] Engler and Renda (n 1).

[212] James Vincent, 'OpenAI Co-Founder on Company's Past Approach to Openly Sharing Research: "We Were Wrong"' *The Verge* (15 March 2023) <https://www.theverge.com/2023/3/15/23640180/openai-gpt-4-launch-closed-research-ilya-sutskever-interview> accessed 24 March 2023.

3. A single point of failure by the developer (e.g. an error in the training data) could create a cascading effect that causes errors for all subsequent downstream users. As European civil society groups have noted: 'A single GPAI system can be used as the foundation for several hundred applied models (e.g. chatbots, ad generation, decision assistants, spambots, translation, etc.) and any failure present in the foundation will be present in the downstream uses.'[213]

In this section, we discuss some of the active, relevant debates in EU and US policy circles around how to regulate GPAI systems, and how the regulation of these systems is further complicated by the dynamics of 'open source' models.

## 6.8 Supply chains and market dynamics for GPAI models

So far, GPAI models have mostly been released on a cloud computing platform and made accessible to other developers via an API but with the capability to fine-tune models using their own data. Many end users will also be likely to access such systems via existing tools, such as operating systems, browsers, voice assistants and productivity software (such as Microsoft Office and Google Workspace).

In the current market structure of cloud computing, Amazon and Microsoft (and to a lesser extent Google's parent company, Alphabet) already have large market shares,[214] with substantial investments into machine learning R&D and global computing and communications infrastructure. It therefore seems likely that these three companies will also become highly successful in offering GPAI models on their platforms. These companies already offer a range of AI services to clients, such as Google's AI Infrastructure and Microsoft's Azure AI Platform. They already "can offer their services at lower cost, broader scale, greater technical sophistication, and with potentially easier access for customers than many competitors."[215] And already,

[213] Access Now et al. (n 17).

[214] Felix Richter, 'Amazon, Microsoft & Google Dominate Cloud Market' (*Statista Infographics*, 23 December 2022) <https://www.statista.com/chart/18819/worldwide-market-share-of-leading-cloud-infrastructure-service-providers> accessed 21 March 2023.

[215] J Cobbe, M Veale and J Singh, 'Moving beyond "Many Hands": Accountability in Algorithmic Supply Chains', *Proceedings of Fairness, Accountability and Transparency '23* (ACM 2023) 9.

“industry concentration is creating toxic competition among AI firms, leading them to release models commercially before they are ready and before they have undergone necessary scrutiny or risk mitigation.”[216]

However, scholars have noted that ’the fact that AIaas operates at scale as an infrastructure service does offer potential points of legal and regulatory intervention. Given AI services will likely be widely used in future, then regulating at this infrastructural level could potentially be an effective way to address some of the potential problems with the growing use of AI’.[217] This would mean focusing regulatory attention on the large providers of these foundational models.

## 6.9 Considerations for assigning responsibility for GPAI models

Drawing on our framework and the principles of efficacy and transparency, it may be more efficient to deal with risks such as bias in suppliers that are higher upstream in supply chains, if their models / systems are being used by large numbers of downstream deployers and developers. Otherwise, ‘excluding [GPAI] models could potentially distort market incentives, leading companies to build and sell GPAI models that minimise their exposure to regulatory obligations, leaving these responsibilities to downstream applications’.[218]

There are concerns that SMEs building systems on top of GPAI models will not have the resources to address many risks. This will present problems because ‘shifting responsibility to these lower-resourced organizations… simultaneously exculpates the actors best placed to mitigate the risks of general purpose systems, and burdens smaller organizations with important duties they lack the resources to fulfil’.[219]

Locating responsibility with GPAI developers higher up the supply chain would enable them to ‘control several levers that might

[216] David Gray Widder, Sarah West and Meredith Whittaker, ‘Open (For Business): Big Tech, Concentrated Power, and the Political Economy of Open AI’ 18 <https://papers.ssrn.com/abstract=4543807> accessed 7 September 2023.

[217] Cobbe and Singh (n 22) 52.

[218] Engler and Renda (n 1) 23.

[219] N Kolt, ‘Algorithmic Black Swans’ (2023) 101 Washington University Law Review 33.

partially prevent malicious use of their AI models. This includes interventions with the input data, the model architecture, review of model outputs, monitoring users during deployment, and post-hoc detection of generated content.' But it will not create a perfect system, rather: 'the efficacy of these efforts should be considered more like content moderation, where even the best systems only prevent some proportion of banned content.'[220]

The US Federal Trade Commission has announced a potentially far-reaching approach under its consumer protection authority, warning businesses creating generative AI systems they should 'consider at the design stage and thereafter the reasonably foreseeable – and often obvious – ways it could be misused for fraud or cause other harm. Then ask yourself whether such risks are high enough that you shouldn't offer the product at all.'[221]

However, as AI software and models become more generalisable and have potentially more users, it becomes harder for their developers to consider customer-specific contexts and potential harms. As scholars have pointed out, 'AI practitioners encounter difficulty in engaging with downstream marginalized groups in large scale deployments. Even where a company is working directly with a client to develop a system for them, it may be 'unable to know what the customer later did with that system after the initial prototype phase, as follow up work does not scale'.[222] Some responsibilities for GPAI supply chains must be placed on deployers who are using the system in a specific context.

Other scholars suggest that systems such as ChatGPT are so general-purpose and usable in so many contexts they should be regulated as a specific category. This would place a duty on developers to actively monitor and reduce risks, in a similar manner to the obligations on platforms of the EU Digital Services Act (Article 34)

[220] Alex Engler, 'Early Thoughts on Regulating Generative AI like ChatGPT' (16 February 2023) <https://www.brookings.edu/blog/techtank/2023/02/16/early-thoughts-on-regulating-generative-ai-like-chatgpt/> accessed 21 February 2023.

[221] Michael Atleson, 'Chatbots, Deepfakes, and Voice Clones: AI Deception for Sale' (*Federal Trade Commission Business Blog*, 20 March 2023) <https://www.ftc.gov/business-guidance/blog/2023/03/chatbots-deepfakes-voice-clones-ai-deception-sale> accessed 22 March 2023.

[222] Widder and Nafus (n 14).

and the UK Online Safety Bill.[223] They also suggest regulators should monitor the 'fairness, quality and adequacy of contractual terms and instructions' between providers and end-users, as is also considered for platforms under the Online Safety Bill.[224] Researchers suggest a specific category of regulation, which imposes limited transparency obligations on generative AI developers, but imposes the duty to implement a risk-management system on companies *using* such a system in high-risk applications.[225]

The European Parliament has proposed tailored requirements for GPAI,[226] 'foundation models'[227] and 'generative AI'.[228] It conceptualises foundation models and generative AI as sub-categories of GPAI, and set different rules for each:

5. GPAI (providers will be required to share information downstream in order to support downstream providers (e.g. fine-tuners) to comply, if deploying the GPAI in a high-risk area.
6. Foundation model providers will have to obligations at the design and development phase, and throughout the lifecycle. The requirements focus on risk and quality management, data governance measures, and testing the model for predictability, interpretability, corrigibility, safety and cybersecurity. These rules are aimed to be "broadly applicable", i.e. independent of distribution channels, modality, or development method.
7. Finally, generative AI providers will be compelled to follow transparency obligations to make clear to end users that they are interacting with an AI model, and will also have to document and make publicly available a summary of the use of training data protected under copyright law.

[223] Michelle Donelan and Lord Parkinson of Whitley Bay, Online Safety Bill 2023. European Parliament and Council of the European Union, 'Digital Services Act' art 34.

[224] Natali Helberger and Nicholas Diakopoulos, 'ChatGPT and the AI Act' (2023) 12 Internet Policy Review <https://policyreview.info/essay/chatgpt-and-ai-act> accessed 22 February 2023.

[225] Philipp Hacker, Andreas Engel and Theresa List, 'Understanding and Regulating ChatGPT, and Other Large Generative AI Models: With input from ChatGPT' (*Verfassungsblog*, 20 January 2023) <https://verfassungsblog.de/chatgpt/> accessed 20 January 2023.

[226] "an AI system that can be used in and adapted to a wide range of applications for which it was not intentionally and specifically designed".

[227] "an AI model that is trained on broad data at scale, is designed for generality of output, and can be adapted to a wide range of distinctive tasks".

[228] Defined as "foundation models specifically intended to be used in AI systems specifically intended to generate, with varying levels of autonomy, content such as complex text, images, audio, or video".

The EU AI Act will therefore regulate GPAI in some form, but the exact requirements will be dependent on negotiations concluding by the end of 2023 or early 2024.

## 6.10 AI system release strategies

One of the biggest factors affecting an AI component's supply chain and how subsequent responsibilities are assigned is how it is released. In some cases, AI components will be released in ways that make downstream developers or deployers incapable of accessing or understanding critical details of how they are trained. In the case of GPAI systems, how a model is released will have significant impacts on how responsibilities for addressing misuse should be applied.

Researchers have summarised various trade-offs for the degree of openness with which developers of 'generative' AI models (those that create new content) make them available to third-parties. More openness can bring benefits, as it increases the ability of a wider range of organisations and experts to audit models, increases the transparency of how models work and brings a broader range of perspectives to bear (while noting 'just because code *can* be audited does not mean that it *will* be').

At the most open end of the spectrum, models released under open-source licences (alongside resources such as training datasets and software) can be developed by communities of developers. This 'fully open' release allows the full details of the model to be made available, which maximises transparency and the opportunity for third-party assessment and development.[229] (Some existing 'open' models use the term as 'more aspiration or marketing than technical descriptor', since 'the term is being applied to widely divergent offerings with little reference to a stable descriptor.')[230]

However, this openness comes with a significant trade-off: reducing the technical ability of developers to constrain their systems' use or misuse. Developers can still implement legal constraints via licences like Responsible AI Licenses (RAIL) that contractually prohibit the use

[229] Irene Solaiman, 'The Gradient of Generative AI Release: Methods and Considerations', *Proceedings of Fairness, Accountability and Transparency '23* (ACM 2023) <http://arxiv.org/abs/2302.04844> accessed 25 February 2023.

[230] Widder, West and Whittaker (n 28).

of the model in a certain way, but it remains unclear how viable this method is as a remedy for preventing misuse.[231] Fully open-source software does not generally impose such limits on deployers, and researchers have noted: 'open source licensing invokes ideological frames that reject the idea that developers should exercise any control at all over harmful use: "the whole point is you can't control that – can't control what people do."'[232]

Applying our framework above, the principles of efficacy and transparency are critical. If a model is released in a more closed manner, it makes it harder for deployers or downstream users in the supply chain to identify these risks. The more closed, the more control a developer has on how a model is designed and used, and therefore the greater the responsibility they should have. The principle of transparency is also critical here, as developers will have far more information than a deployer about the model's architecture. Without transparency mechanisms in place, it will be hard for downstream deployers to identify or mitigate risks.

## 6.11 The challenges of open-source

Open-source GPAI projects play two key roles:

- 'they disseminate power over the direction of AI away from well-resourced technology companies to a more diverse group of stakeholders.
- 'they enable critical research, and thus public knowledge, on the function and limitations of GPAI models.'[233]

While it may seem in the financial interest of companies investing heavily in the development of proprietary models to control their availability, even the largest technology firms are also contributing to open-source systems. For example, Microsoft has contributed to research leading to improvements in the Stable Diffusion image

[231] Danish Contractor and others, 'Behavioral Use Licensing for Responsible AI', *2022 ACM Conference on Fairness, Accountability, and Transparency* (ACM 2022) <https://dl.acm.org/doi/10.1145/3531146.3533143> accessed 24 March 2023.

[232] Widder and Nafus (n 14).

[233] Alex Engler, 'The EU's Attempt to Regulate Open-Source AI Is Counterproductive' (*Brookings Institution TechTank*, 24 August 2022) <https://www.brookings.edu/blog/techtank/2022/08/24/the-eus-attempt-to-regulate-open-source-ai-is-counterproductive/>.

generation system.[234] However, it is likely such contributions will be in the interests of the companies concerned.[235] An expert review for the European Commission found that platforms generally shape innovation within their own ecosystems to bolster their business models,[236] while companies making their models available under open (to some extent) licences are easily able to incorporate improvements made by other developers directly back into their products.[237]

It is not yet clear whether the very high resource requirements of creating the highest-capability models (such as OpenAI's GPT-4 and Google's LaMDA) will mean regulating their safety and availability via those companies will be feasible (as called for by OpenAI's CEO[238] and others).

While open-source generative language models have been advancing at a rapid pace, so far they have been significantly based on models from firms such as Meta, whose LLaMA was leaked in March 2023[239] and which is now made available under licences with significant restrictions and lacking important information.[240] The AI Now Institute suggests: "Even if costs are lower or come down as these systems are deployed at scale (and this is a hotly contested claim), Big Tech is likely to retain a first mover advantage".[241] While fine-tuning these models for specific applications is much less computationally expensive than first creating them, "the fine-tuned end products largely function as barnacles on the hull of Big Tech, rather than a meaningful alternative to it. They still need to be run on Big Tech infrastructures (as a rule), and cede power to define and create the

[234] Yuheng Li and others, 'GLIGEN: Open-Set Grounded Text-to-Image Generation' (17 April 2023) <http://arxiv.org/abs/2301.07093> accessed 6 March 2023.

[235] Meredith Whittaker, 'The Steep Cost of Capture' (2021) 28 Interactions 50.

[236] Ariel Ezrachi and Maurice E Stucke, 'Digitalisation and Its Impact on Innovation' (European Commission DG Research and Innovation 2020) 978-92-76-17462–2, KI-BD-20-003-EN-N <https://research-and-innovation.ec.europa.eu/knowledge-publications-tools-and-data/publications/all-publications/digitalisation-and-its-impact-innovation_en> accessed 21 March 2023.

[237] Widder, West and Whittaker (n 28) 11-12.

[238] 'Oversight of A.I.: Rules for Artificial Intelligence' <https://www.judiciary.senate.gov/download/2023-05-16-testimony-altman> accessed 21 May 2023.

[239] D Patel and A Ahmad, 'Google "We Have No Moat, And Neither Does OpenAI"' (*semianalysis*, 4 May 2023) <https://www.semianalysis.com/p/google-we-have-no-moat-and-neither> accessed 19 May 2023.

[240] Widder, West and Whittaker (n 28).

[241] Amba Kak and Sarah Myers West, 'AI Now 2023 Landscape: Confronting Tech Power' (AI Now Institute 2023) 17 <https://ainowinstitute.org/2023-landscape> accessed 21 May 2023.

core model logics to the large companies who have the resources to create them from scratch."[242]

While it would be possible for legislation to go further in applying obligations to online distribution of open-source AI components, its likely efficacy would be severely open to question, given the following observations:

- Without comprehensive international agreement (which is difficult to imagine in the current geopolitical climate), unrestricted development and sharing would be likely to continue in other jurisdictions (including the USA, whose constitution includes strict restrictions on government limits on publication).[243]
- The underlying techniques and data used for training models are likely to continue circulating freely (open-source software and a public molecule database were used to train a model used to identify potential biochemical weapons[244] which received significant media attention).
- Such restrictions would be likely to significantly impede the pace of research and development relating to AI tools and techniques, including those to identify and remedy potential harms, particularly outside of the large firms which already and increasingly dominate AI research.[245]

While not a precise analogy (because large AI models are much more complex and resource-intensive to create than encryption software), attempts by the USA and its allies to control the global spread of encryption technology throughout the 1980s and 1990s ultimately failed for similar reasons.[246]

Recognising this, advocates for regulation of 'frontier' (most capable) AI systems have suggested using controls on the sale of specialised

[242] Widder, West and Whittaker (n 28) 18.

[243] Andrea Matwyshyn, 'Hacking Speech: Informational Speech and the First Amendment' (2013) 107 Northwestern University Law Review 795.

[244] Fabio Urbina and others, 'Dual Use of Artificial-Intelligence-Powered Drug Discovery' [2022] Nature Machine Intelligence 189.

[245] Nur Ahmed, Muntasir Wahed and Neil C Thompson, 'The Growing Influence of Industry in AI Research' (2023) 379 Science 884.

[246] Whit Diffie and Susan Landau, *Privacy on the Line* (Updated and Expanded Edition, Random House 2010) <https://www.penguinrandomhouse.com/books/654750/privacy-on-the-line-updated-and-expanded-edition-by-whitfield-diffie-and-susan-landau/> accessed 12 March 2023.

processors (from companies such as Nvidia) necessary to create these models to enforce 'safe and ethical uses of the technology'.[247] But there is no sign of such political action in the US, where it would likely have most effect. A deeply divided Congress is unlikely to agree such laws in the medium term, given both strong political disagreement on what 'safe and ethical' uses of technology looks like, and the potential impact on US company profits and global technical leadership.

## 6.12 Conclusion

Any approach to AI regulation will need to grapple with different supply chains behind those services and with assigning responsibilities to actors in those supply chains. Broadly speaking, policymakers and regulators will need to understand "who is doing what for whom, who is performing what key functions for others, who is core to certain supply chains, and who is systemically important."[248]

Transparency mechanisms like model cards, datasheets, etc. are an essential component of supply chain accountability, but can come into tension with other incentives, such as trade secrecy. OpenAI's recent release of GPT-4 and Google's recent release of Bard saw both companies refuse to provide details on the models' architecture and data sources, citing reasons of competition and safety.[249]

The refusal by companies to make these details accessible should alarm regulators and policymakers, as it removes the ability of downstream users and third-party auditors to assess the safety, performance and ethical considerations of these models. These transparency mechanisms should be standardised by governments and regulators, ideally via international standards and requirements, and made a legal requirement from companies putting AI models and services on the market.

[247] Richard Waters, 'US Should Use Chip Leadership to Enforce AI Standards, Says Mustafa Suleyman' Financial Times (1 September 2023) <https://www.ft.com/content/f828fef3-862c-4022-99d0-41efbc73db80> accessed 8 September 2023.

[248] Cobbe, Veale and Singh (n 27) 12.

[249] James Vincent, 'OpenAI Co-Founder on Company's Past Approach to Openly Sharing Research: "We Were Wrong"' *The Verge* (15 March 2023).

Where AI components are used by many downstream companies in a supply chain, it will be more efficient for some issues to be fixed by the component developer. Allocation of responsibility must also account for the power imbalances between different actors and how AI systems are released. Those developing an AI system may be in a greater position of power over their suppliers or users to contractually offload responsibilities. Depending on how an AI system is released, upstream providers may need to bear more responsibility to evaluate and address the potential issues within their system.

General-purpose AI (GPAI) systems complicate supply chain considerations. Determining what kinds of responsibilities should apply will require both *ex-ante* assessments of risk and assignments of responsibility by regulators and policymakers, along with *ex-post* regulation of the actual uses of these systems.

As with other digital markets such as search, social networking services and especially cloud computing, competition concerns are likely to arise in the provision of AI services, due to high returns to scale and the importance of access to specific data, compute and labour resources.[250]

Open-source technologies further complicate supply chain considerations. Regulation must address how AI technologies (and powerful components of those AI technologies, like underlying models, datasets or model weights) are released. But there are strong practical benefits for innovation, public accountability and competition from the availability of open-source tools. Limits on publication of components to manage risks face significant constraints, not least the small probability of the international agreement which would be needed to make them remotely effective, and the freedom of expression implications of trying to limit access to the underlying knowledge.

[250] Widder, West and Whittaker (n 28) 7-11.

# 7 GenAI and the Goblet of Compliance: Delving into the Pensieve of Privacy Principles

**Shruti Shreya, Graduate Student, O.P. Jindal Global University, India;**

**Pranav Bhaskar Tiwari, Graduate Student, O.P. Jindal Global University, India;**

**Gyan Prakash Tripathi, Advocate, Delhi High Court, India**

## Abstract

The intersection of AI and human society necessitates robust regulatory frameworks. With the emergence of ChatGPT in 2022, the EU AI Act led the charge in governing technologies like generative AI (GenAI). As GenAI integrates across sectors, it presents distinct challenges, from potential breaches of business confidentiality to concerns of academic integrity. This working paper represents the first step in a three-phase research initiative, centred on the development of a comprehensive privacy compliance framework for GenAI. Through careful legal analysis and engagement with stakeholders, we establish sixteen key privacy principles tailored for GenAI platforms. The ensuing stages aim to refine this framework based on broad stakeholder feedback and test the framework's applicability across various GenAI platforms, ensuring users' privacy rights remain paramount. This research offers both a timely insight into GenAI's evolving legal landscape and a blueprint for future studies and regulatory endeavours.

## 7.1 Introduction

The interplay between artificial intelligence (AI) and humanity has unfolded over several decades. Institutions have progressively formulated ethical guidelines to govern their interaction with AI technologies.[251] In the past decade, governments worldwide have intensified their scrutiny of the AI landscape, driven by the impetus to

[251] The Future of Life Institute. Asilomar AI Principles. (2017). Retrieved from <https://futureoflife.org/ai-principles/>. See also: Organisation for Economic Co-operation and Development (OECD). (2022). Classification of artificial intelligence: A two-pager. Retrieved from <https://wp.oecd.ai/app/uploads/2022/02/Classification-2-pager-1.pdf>. Montreal Declaration for Responsible Development of Artificial Intelligence. (2018). <https://recherche.umontreal.ca/english/strategic-initiatives/montreal-declaration-for-a-responsible-ai/>.

establish regulatory frameworks. Notably, the EU AI Act has assumed a pioneering role in delineating comprehensive regulations that encompass even software.[252] Within this landscape, the emergence of generative AI (GenAI) has exerted seismic shifts in user interactions with the Internet, exemplified by the public debut of ChatGPT in November 2022.[253]

GenAI has permeated diverse spheres of human existence be it business, education, or recreation.[254] However, its transformative potential has engendered apprehensions among various stakeholders. Businesses harbour concerns over the prospect of employees inadvertently divulging proprietary company information into the enigmatic "Blackbox" of GenAI.[255] This information could potentially be harnessed for training purposes by the platform, prompting apprehensions within the business community.[256] Concurrently, labour commissions express concerns about its ramifications on employment dynamics.[257] Educational institutions, in turn, harbour reservations regarding the preservation of academic integrity, given GenAI's influence on student submissions.[258]

In the midst of this intricate panorama lies an opportunity not solely to bridge the digital divide but also to cultivate a more inclusive digital

[252] European Union. (2021). Proposal for a regulation of the European Parliament and of the Council on artificial intelligence (Artificial Intelligence Act). Retrieved from <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A52021PC0206>; *See also* Responsible AI Institute. (2022, May 4). A framework to navigate the emerging regulatory landscape for AI. OECD AI Policy Observatory. Retrieved from <https://oecd.ai/en/wonk/emerging-regulatory-landscape-ai>.

[253] OpenAI. (2022, February 23). ChatGPT: Generative pre-trained transformer for conversational applications. Retrieved from <https://openai.com/blog/chatgpt/>.

[254] Walsh, M., & Veale, M. (2022). Generative AI in art and design. *AI & Society*, 37(1), 1-18. *See also* Danks, D., & Nielsen, M. (2021). Generative AI in finance. *Journal of Financial Stability*, 43, 100958; Danaher, J., & Devlin, K. (2020). Generative AI in healthcare. *Nature Medicine*, 26(10), 1355-1357; Dietterich, T., & Hohman, M. (2019). Generative AI in manufacturing. *Manufacturing Letters*, 18, 1-5; Wardrip-Fruin, N., & Mateas, M. (2018). Generative AI in media and entertainment. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 25(6), 1-27.

[255] Susarla, A., Chui, M., & Osborne, M. (2020, March). The black box of AI: How to mitigate the risks of unexplained bias. *Harvard Business Review*.

[256] Veale, M., & Walsh, M. (2022). The copyright challenges of generative AI. *AI & Society*, 37(1), 1-18. See also: Calo, R., & Buccafusco, C. (2020). Generative AI and the future of copyright. The Yale Law Journal, 130(1), 1-60.

[257] World Economic Forum. (2020). The Future of Jobs: Jobs and Skills in 2030. Retrieved from <https://www.weforum.org/reports/the-future-of-jobs-report-2020/>.

[258] University of Melbourne. (2023, March 8). Advice for students regarding Turnitin and AI writing detection. Retrieved from <https://academicintegrity.unimelb.edu.au/plagiarism-and-collusion/artificial-intelligence-tools-and-technologies/advice-for-students-regarding-turnitin-and-ai-writing-detection>.

landscape, particularly for differently-abled members of our society.[259] The societal expectations vested in GenAI are undeniably substantial. Notably, the proliferation of such platforms has not followed a linear trajectory, but rather an exponential one.[260] It is pivotal to recognise that many GenAI platforms are constructed upon publicly available, and at times intellectually protected, information.[261] Their engagements across diverse societal segments necessitate rigorous alignment with overarching privacy norms, safeguarding users' fundamental right to privacy.

This tripartite research endeavour to comprehensively comprehend and subsequently influence the guiding tenets underpinning GenAI development. The primary focus of this paper, constituting the first stage, involves the systematic construction of a privacy compliance framework tailored for GenAI platforms. This entails a doctrinal examination of existing jurisprudence, deftly contextualised for GenAI. The subsequent stages encompass a multifaceted approach. The second stage entails active engagement with stakeholders across the ecosystem, soliciting essential insights through semi-structured interviews and focused group discussions, thus enriching the framework's contours. The concluding stage culminates in an empirical assessment of the adherence exhibited by a representative set of GenAI platforms to the developed framework. In due course, this proposed framework stands poised to empirically ascertain the progressive evolution of platforms in enhancing their compliance posture over time.

In this paper, we begin by delineating the research methodology employed for this doctrinal study. The subsequent section sheds light on the foundation of the privacy principles integrated into the framework, with their selection being contextualised based on

[259] Erhardt, J., & Krishnan, V. (2023, August). Designing generative AI to work for people with disabilities. *Harvard Business Review*. <https://hbr.org/2023/08/designing-generative-ai-to-work-for-people-with-disabilities>.

[260] Grand View Research. (2023). The rise of generative AI platforms: A market research report. Grand View Research.

[261] Berrada, M., Jacovides, A., & Ouedraogo, A. (2023, February). Navigating intellectual property rights in the era of generative AI: The crucial role of educating judicial actors. *UNESCO*. <https://www.unesco.org/en/articles/navigating-intellectual-property-rights-era-generative-ai-crucial-role-educating-judicial-actors?TSPD_101_R0=080713870fab2000186b2cce03d54213b1bdc1c26f57b6eea3d762f6edc2b3144e3067cff8>.

the socio-technical systems theory.[262] This is achieved through a rigorous examination of global privacy scholarship. Moreover, we delved into academic works to pinpoint the juncture at which the identified privacy principles are evaluated in the life cycle of GenAI platforms. Our study concludes by assessing the privacy compliance landscape for GenAI platforms, deliberating on subsequent steps for procuring feedback on the framework, and spotlighting the identified research lacunae in the GenAI privacy ambit. Additionally, an annexure is appended, furnishing the privacy compliance framework for prospective adoption by researchers or organisations.

## 7.2 Research Methodology

### 7.2.1 Objective

This research aims to develop and validate a privacy compliance framework for GenAI platforms. The methodology is structured into three interconnected stages, with the current paper set to focus on the first stage.

### 7.2.2 Relevance

GenAI has evolved from a niche technological tool to a widely accessible platform integral to daily life. Unlike other AI forms, GenAI's significant human interaction, owing to its user-friendly interface, means it is not just for the technical elite but the everyday user. This widespread adoption, coupled with its rapid market growth, necessitates stringent regulatory oversight.

The EU AI Act's Chapter 3 underscores this by categorising GenAI as a high-risk AI system, highlighting the pressing need for tailored regulations.[263] Given GenAI's unique attributes and its profound societal implications, a dedicated privacy compliance framework is imperative to ensure its responsible evolution.

[262] Caraher, T. P., & Anderson, R. J. (2017). Socio-technical systems theory. In Business. Leeds. ac.uk. Retrieved from <https://business.leeds.ac.uk/research-stc/doc/socio-technical-systems-theory#:~:text=Socio%2Dtechnical%20theory%20has%20at,parts%20of%20a%20complex%20systems>.

[263] Chapter 3 European Union. (2022). Regulation (EU) 2022/765 of the European Parliament and of the Council on artificial intelligence (Artificial Intelligence Act). Official Journal of the European Union.

## 7.3 Research Design

### Stage 1: Literature Analysis

- **Scope:** A systematic review of existing privacy scholarship, legislative instruments, and policy documents undertaken to identify relevant privacy principles for GenAI platforms.
- **Outcome:** A foundational comprehension of privacy principles and their prospective application to GenAI platforms will be achieved, leading to the formulation of a draft privacy compliance framework.

### Stage 2: Stakeholder Engagement

- **Scope:** Hosting semi-structured interviews and focus group discussions to seek inputs on the draft privacy compliance framework. The possibility of anonymous feedback mechanisms will also be explored. This stage will engage technical experts, legal professionals, AI researchers, sociologists, organisational leaders, civil society and users among others.[264]
- **Outcome:** Feedback will guide the refinement of the initial framework, ensuring its practicality and applicability.

### Stage 3: Empirical Analysis

- **Scope:** An primary analysis will be undertaken on a select set of GenAI platforms, gauging their alignment with the updated framework. Expert insights from stage 2 will guide the platform selection. The criteria for 'compliance' and 'non-compliance' will be explicitly set, drawing from initial research and stakeholder insights.
- **Outcome:** The analysis will yield insights into privacy compliance by platforms, spotlighting areas of alignment and divergence.

### 4. Research Questions

The inaugural stage of this research seeks to address:

- Which foundational privacy principles, drawn from contemporary legal and technical sources, will be pertinent to GenAI platforms?
- During which phase of the GenAI lifecycle should these principles be most effectively implemented?

[264] Veale, M., & Brass, A. (2020). The stakeholders of artificial intelligence: A literature review. *Ethics and Information Technology*, 22(1), 1-18.

### 5. Methodological Considerations

- **Doctrinal Approach:** The first stage involves a doctrinal research approach, investigating global privacy law resources, inclusive of statutes, regulations, case laws, and scholarly discourses. Policies and directions from global corporate and governmental entities will also be reviewed.
- **Qualitative Data Analysis:** Qualitative analysis tools will be employed for data sourced from stakeholder interactions, ensuring consistent and objective interpretation in stage 2.
- **Ethical Concerns:** Given the nuanced overlap of GenAI and privacy, attention will be devoted to the ethical ramifications of the findings. Data sourcing, stakeholder feedback, and GenAI's broader societal impacts will be duly considered.
- **Comparative Analysis:** A section will be integrated that contrasts the privacy stances of varied global territories towards GenAI, enriching the research with diverse global insights.

### 6. Limitations and Challenges

This paper is aimed towards crafting a comprehensive privacy compliance framework. The paper, currently in stage 1, is based on doctrinal research. Refinements will be anchored on stakeholder feedback, ear-marked for stage 2, ensuring that the framework remains agile and attuned to the shifting terrains of GenAI.

## 7.4 Deploying the Foundational Privacy Principles for Gen AI Platforms

### 7.4.1 Identifying the stages of the AI lifecycle

The efficacy of privacy principles in evaluating the privacy-readiness of GenAI platforms hinges upon their application in the various stages of the lifecycle of the model powering them. After meticulous scrutiny, we found Silva and Alahakoom's framework to be both exhaustive and germane for our research.[265] The scholars delineated the primary stages – design, develop, and deploy – and further segmented these into 19 nuanced sub-stages. Such a comprehensive

[265] De Silva, D., & Alahakoon, D. (2022). An artificial intelligence life cycle: From conception to production. Patterns, 3(6), 100489.

partition ensures that the myriad use-cases of GenAI are encapsulated within the lifecycle. Their “CDAC AI life cycle” (CDAC framework) has demonstrated the need for a life cycle approach that has been conceived exclusively to address the challenges of designing, developing, deploying, and managing an AI solution. They have also attempted to address the drawbacks in previous frameworks and further enable continuous, multi-granular expansion of the overarching preliminary risk assessment through its constituent stages and phases.

In the context of our proposed AI pipeline, the design stage encompasses aspects like data collection, annotation, documentation, addressing social, ethical, and cultural factors, implementing best data practices, recording consent processes, exploring design methods, evaluating explainability and interpretability, and fostering human-AI collaboration. These align closely with the design phase of the CDAC framework. Similarly, the development phase of the CDAC framework has guided our identified principles of equality, safety, responsibility, inclusivity, non-discrimination, transparency, accountability, privacy, and security. Lastly, the AI model operationalisation and deployment encompassing explainability, responsibility, accountability, data documentation practices, evaluation, and monitoring, is mirrored in aspects of the CDAC framewrok’s deploy phase.

To fortify this model, recognising the high-risk potential of GenAI platforms, we introduced an additional ‘audit’ layer. In instituting this layer, we drew upon the insights of Haakman et al.[266] Their seminal work underscores the importance of documentation, model monitoring, and model risk assessment, thereby steering AI models towards greater explainability, accountability, and oversight. This enhanced focus augments risk assessment capabilities. With this working paper, we extend an invitation for expert feedback, aspiring to refine the AI lifecycle stages further, ensuring robust and responsible AI deployment.

[266] Haakman, M., Cruz, L., Huijgens, H., & van Deursen, A. (2021). AI lifecycle models need to be revised: An exploratory study in Fintech. Empirical Software Engineering, 26(5), 1-29.

## 7.5 Identifying the Privacy Principles for the GenAI Platforms

In the swiftly evolving domain of GenAI, the imperative of privacy is clear and pressing. This chapter endeavours to elucidate a set of sixteen principles, each one of them paramount in guiding and shaping the deployment of GenAI platforms. Drawing inspiration from the sociotechnical systems theory,[267] which underscores the interplay between technological advancements and societal structures, we recognise that effective privacy measures for GenAI cannot be solely about design or mere technical specifications. Rather, they must take into account the broader societal contexts in which these technologies operate. Each principle, thus, is not just a technical directive but resonates with its broader implications on societal, ethical, and legal dimensions. As we delve into each, the lens of sociotechnical systems theory aids in understanding their significance beyond the confines of technology, grounding them in the lived realities of individuals and communities. In keeping with this holistic perspective, our focus remains both on the theoretical foundations and their pragmatic applications in the real-world context of GenAI. It is pertinent to highlight that we initiate the analysis with the globally recognised work of Dr. Ann Cavoukian on privacy by design (PbD) principles[268] that incorporate multiple principles within its ambit and then build on to more niche aspects like explainability.

### i. PbD1: Proactive not Reactive; Preventative not Remedial

At the very core of the GenAI lifecycle – encompassing design, development, and deployment stages – lies the principle of proactive forethought in privacy management.[269] By embedding privacy measures at the foundational level, GenAI platforms can eschew the pitfalls of retroactive amendments. For example, during a GenAI

[267] Caraher, T. P., & Anderson, R. J. (2017). Socio-technical systems theory. In Business. Leeds. ac.uk. Retrieved from <https://business.leeds.ac.uk/research-stc/doc/socio-technical-systems-theory#:~:text=Socio%2Dtechnical%20theory%20has%20at,parts%20of%20a%20complex%20systems>.

[268] Cavoukian, A. (2009). Privacy by design: The seven foundational principles. Information & Privacy Commissioner of Ontario. Retrieved from <https://www.ipc.on.ca/wp-content/uploads/resources/7foundationalprinciples.pdf>.

[269] Cavoukian, A. (2009). Privacy by design: The seven foundational principles. Information & Privacy Commissioner of Ontario. Retrieved from <https://www.ipc.on.ca/wp-content/uploads/resources/7foundationalprinciples.pdf>.

model's training phase, proactive strategies can ensure adherence to privacy-respecting norms, thereby diminishing potential breaches upon deployment. A tangible example of such foresight would be a foundational model refraining to train on publicly accessible personal data.[270]

### ii. PbD2: Privacy as the Default Setting

Generative AI platforms, given their multifaceted nature and vast data processing capabilities, necessitate privacy as a default across design, deployment, and auditing stages. Such an approach ensures automatic privacy safeguarding without continual human oversight. In concrete terms, this would necessitate that GenAI models are hardwired to exclude private data markers unless an explicit user permission is obtained.[271] Drawing parallels, Apple's Siri epitomises this with its on-device personalisation, minimising data transfers to external servers and consequently bolstering user privacy.[272] The optimal pathway for GenAI platforms would ensure any learning derived from user interactions remains on-device by default, unless informed user consent dictates otherwise, as illustrated by ChatGPT's pivot post the Italian data protection authority's mandate.[273]

### iii. PbD3: Privacy Embedded into Design

For GenAI platforms to truly internalise the 'Privacy Embedded into Design' principle, a proactive stance is imperative throughout design and development.[274] This includes exhaustive Privacy Impact Assessments,[275] the adoption of a privacy-centric architectural blueprint,[276] immediate

[270] The Guardian. (2023, April 10). 'I didn't give permission': Do AI's backers care about data law breaches? Retrieved from <https://www.theguardian.com/technology/2023/apr/10/i-didnt-give-permission-do-ais-backers-care-about-data-law-breaches>.

[271] Artificial Intelligence/Machine Learning Risk & Security Working Group (AIRS). Artificial Intelligence Risk & Governance. Wharton School of the University of Pennsylvania. Retrieved from <https://aiab.wharton.upenn.edu/research/artificial-intelligence-risk-governance/>.

[272] Apple. (2023, January 27). How Siri Works. Retrieved from https://support.apple.com/en-us/HT20704>.

[273] Gaudiosi, J. (2023, April 1, 2023). ChatGPT is once again available in Italy after a temporary ban. *Engadget*. Retrieved from <https://www.engadget.com/chatgpt-is-once-again-available-in-italy-after-a-temporary-ban-195716663.html>.

[274] Cavoukian, A. (2009). Privacy by design: The seven foundational principles. Information & Privacy Commissioner of Ontario. Retrieved from <https://www.ipc.on.ca/wp-content/uploads/resources/7foundationalprinciples.pdf>.

[275] Barocas, S., Hardt, M., Narayanan, A., & Selbst, A. D. (2019). The ethics of artificial intelligence: Mapping the debate. *Nature*, *569*(7755), 525-531.

[276] Fung, A., Yu, H., & Wright, J. (2021). Privacy-preserving artificial intelligence: A survey. *ACM*

data anonymisation post-collection,[277] and data acquisition limitations to essential elements only.[278] From a development standpoint, adherence to secure development conventions, establishment of stringent data management norms, utilisation of transparent algorithms, and incorporation of user-centric privacy features are non-negotiable.[279] Consistent vulnerability assessments and an unwavering focus on privacy-driven training and thorough documentation are cornerstones of this approach. These methodologies ensure privacy remains a constitutive aspect of GenAI systems, bolstering user trust.

### iv. PbD4: Full Functionality — Positive-Sum, not Zero-Sum

GenAI platforms often grapple with the dichotomy of amplifying privacy measures and maintaining peak functionality, evident in the intricacies of implementing differential privacy,[280] the convolution of extensive privacy controls, and latency introduced by real-time data anonymisation. Yet, the 'Full Functionality — Positive-Sum, not Zero-Sum principle challenges this duality,[281] heralding innovative techniques that simultaneously respect both facets. For instance, hybrid models merging synthetic and differentially private data can maintain model efficacy, interfaces can feature tiered privacy controls, and latency issues might find mitigation through enhanced algorithms or edge computing. Such insights reinforce the notion that GenAI platforms, while complying with regulatory edicts, need not trade off robust privacy measures for core functionality.

### v. PbD5: End-to-End Security — Full Lifecycle Protection

Embracing an end-to-end perspective, GenAI platforms ought to champion privacy-centric architecture right from inception.[282] This calls

*Computing Surveys*, *54*(4), 1-37.

[277] Gupta, A., & Das, A. K. (2022). A survey on data anonymization techniques for general-purpose artificial intelligence systems. *ACM Computing Surveys*, *55*(2), 1-41.

[278] Erlingsson, Ú., Kantarcioglu, M., & Zhang, L. (2020). Privacy-preserving personalization in voice assistants. *ACM Transactions on Information Systems Security (TISSEC)*, *23*(3), 1-33.

[279] Barocas, S., & Selbst, A. D. (2016). The future of artificial intelligence and privacy. *Harvard Law Review*, *131*(1), 193-238.

[280] Ghosh, A., & Kantarcioglu, M. (2020). Differential privacy in generative AI: A survey. *ACM Computing Surveys*, *53*(1), 1-38. See also: Dwork, C. (2006). Differential privacy. *Automated Decision Making*, 1(2), 23-40.

[281] Cavoukian, A. (2009). Privacy by design: The seven foundational principles. Information & Privacy Commissioner of Ontario. Retrieved from <https://www.ipc.on.ca/wp-content/uploads/resources/7foundationalprinciples.pdf>.

[282] Bhargava, H. K., Kantarcioglu, M., & Zhang, L. (2020). Privacy-centric architectural blueprint for artificial intelligence systems. *IEEE Security & Privacy*, *18*(5), 68-75.

for meticulous field-level encryption of sensitive data[283] and rigorous role-based access control mechanisms during development.[284] Periodic external security audits, paired with transparent data retention and erasure protocols, ensure continued adherence to privacy norms.[285] Additionally, the existence of a predetermined incident response plan, in line with legal prerequisites, promises timely interventions during data breaches. This panoramic strategy ensures the maintenance of data privacy and security across the GenAI lifecycle.

### vi. PbD6: Visibility and Transparency — Keep it Open

GenAI platforms should remain committed to presenting user-centric Privacy Impact Assessments, outlining the data's entire journey, from acquisition and processing to storage.[286] Rigorous audit logs, which provide users with a detailed account of data interactions, become crucial. A yearly transparency report, encompassing data interactions, breaches, and consequent rectifications, becomes a testament to the platform's dedication to openness.[287] Scholars like Solove have underscored that institutions collecting data should be transparent about their practices and held accountable for breaches and misuse.[288] Furthermore, fostering avenues for user feedback, coupled with swift and legislatively compliant responses, cultivates a culture of transparent dialogue, balancing transparency with competitive edge.

### vii. PbD7: Respect for User Privacy — Keep it User-Centric

Prioritising user privacy, GenAI platforms must ensure clarity and transparency in consent mechanisms. For instance, a GenAI health tool should proffer a lucid and succinct consent form before accessing health records. Privacy settings, akin to easily navigable dashboards in a GenAI photo utility, should grant users unobstructed control

[283] Li, X., Zhang, L., Kantarcioglu, M., & Choo, K.-K. R. (2017). Field-level encryption: A survey. *ACM Computing Surveys*, *49*(4), 1-35.

[284] Barth, A., Volkamer, M., & Sadeh, N. (2020). Privacy-preserving generative AI: A survey on mechanisms and challenges. *ACM Computing Surveys*, *53*(1), 1-38.

[285] Bartoli, A., De Matteis, S., & Maggi, F. (2022). Privacy-preserving data retention and erasure in generative AI platforms: A survey. *ACM Computing Surveys*, 55(1), 1-36.

[286] Michalsons. (2023, February 15). Privacy impact assessments for generative AI. *Michalsons*. Retrieved from <https://www.michalsons.com/blog/privacy-impact-assessments-for-generative-ai/65772>.

[287] Zhang, L., Barth, A., & Volkamer, M. (2022). Transparency reports for generative AI platforms: A review and research agenda. *ACM Computing Surveys*, *55*(1), 1-36.

[288] Solove, D. J. (2008). Understanding privacy. *Harvard Law Review*, *125*(3), 421-549.

over their data. It is equally vital to recast privacy policies, veering away from dense treatises to crisp, clear documents, facilitating user comprehension without inundation.[289] Through such measures, GenAI platforms can underscore their unwavering allegiance to user-focused privacy in the design, development and deployment stage of the GenAI lifecycle fostering user empowerment.[290] In a notable instance, OpenAI, under the direction of the Italian data protection authority, introduced modifications to its privacy practices.[291] This move, geared towards adherence to the EU's privacy regulations, saw OpenAI integrating enhanced privacy disclosures and affording EU users with more explicit controls over their personal data. These adjustments highlight the importance for GenAI platforms to constantly evolve their privacy measures to ensure they align with user-focused privacy throughout the GenAI lifecycle.

### viii. Notice

Transparency in GenAI platforms' data handling pivots on the "notice" principle, compelling platforms to elucidate their modus operandi regarding data collection, utilisation, and dissemination.[292] Initial interactions could introduce succinct data collection notifications, supplemented by periodic reminders and user-centric dashboards, providing a panoramic view of their data's journey. The introduction of dynamic content warnings—particularly for sensitive data—and instantaneous alerts during third-party sharing can amplify user confidence. A paramount consideration is the demystification of policy documents, ensuring they are intelligible, free from convoluted terminology, and therefore inescapably transparent.[293] Such practices enable GenAI platforms to judiciously fulfil legal requirements while concurrently buttressing user trust.

[289] Privacy in Context: Technology, Policy, and the Integrity of Social Life" (2010). <https://hci.stanford.edu/courses/cs047n/readings/Privacy_in_Context.pdf>.

[290] Cohen, J. E. (2012). Configuring the networked self: Law, code, and the play of everyday practice. Yale University Press.

[291] Garante per la protezione dei dati personali. (2023, February 23). Retrieved from <https://www.garanteprivacy.it/web/garante-privacy-en/home_en>.

[292] GDPR. (2016). Information to be provided where personal data are collected from the data subject. Article 13. General Data Protection Regulation. Retrieved from <https://gdpr-info.eu/art-13-gdpr/>.

[293] European Union Agency for Network and Information Security. (n.d.). Privacy notice. Retrieved from <https://gdpr.eu/privacy-notice/>.

### ix. Data Minimisation

In embracing the principle of data minimisation,[294] GenAI platforms ought to champion a precision-focused approach to data collection[295]—meticulously gathering only pertinent data and discarding the extraneous. Regular audits could act as gatekeepers, ensuring data's continued relevance, whilst compliance modules echoing GDPR principles further reinforce the commitment. Transparent user dashboards enhance trust, providing users a lens to scrutinise and control the data cache. Deploying advanced storage strategies and bias scrutiny mechanisms can enrich data quality, positioning both efficiency and ethical principles at the helm. This meticulous approach accentuates user privacy and adeptly steers through regulatory mazes.

### x. Purpose Limitation

GenAI platforms' allegiance to the "Purpose Limitation" principle[296] necessitates a stringent confinement of data usage to its preordained intent.[297] This not only optimises data veracity and model performance but is emblematic of an ethical data culture, discouraging superfluous data explorations. Further, the proposed EU AI Act mandates that data utilised for training and operating the GenAI system should be in strict adherence to the GDPR.[298] This encompasses the lawful and equitable collection of data, ensuring its use is consistent with the initial collection purpose and upholding its security. By demarcating and abiding by these data use parameters, platforms diminish breach vulnerabilities and cultivate a lucid bond with users, succinctly demystifying the data collection rationale.[299]

[294] Article 5(1)(c) of the General Data Protection Regulation. (n.d.). Retrieved from <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32016R0679>, Article 4(1)(c) of Regulation (EU) 2018/1725. (n.d.). Retrieved from <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32018R1725>.

[295] Martens, M., Rau, M., & Scherrer, S. (2018). Data minimization in the age of big data: A review of concepts, methods, and tools. *Computer Law & Security Review*, 34(1), 107-124.

[296] Article 5 of the General Data Protection Regulation. (n.d.). Retrieved from <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32016R0679>.

[297] Bygrave, P. (2014). The principle of purpose limitation. In P. De Hert, & P. De Wever (Eds.), *Data protection law: A practical guide* (pp. 33-48). Cambridge University Press.

[298] The artificial intelligence act: A new regulation for artificial intelligence in the European Union. Retrieved from https://www.europarl.europa.eu/RegData/etudes/STUD/2020/641530/EPRS_STU(2020)641530_EN.pdf>.

[299] Cavoukian, A., & Singh, S. (2020). The importance of purpose limitation in the age of Gen AI. *Computer Law & Security Review*, 36(4), 1021-1034.

### xi. Right to Erasure

The "Right to Erasure" is an indispensable tenet within GenAI platforms.[300] It magnifies user trust, bequeathing data sovereignty to individuals, thus potentially invigorating platform participation.[301] This right further amplifies data minimisation values, neutralising data breach threats. Taking a leaf out of Google's playbook, OpenAI accords its European users the privilege to challenge the processing of their personal data.[302] AI training is complex: though the model does not store personal data, it still retains traces of its training set, making it hard to remove individual marks entirely. To address this, it is crucial to retrain models without the data of specific users. This manoeuvre not only navigates the technical minefields but also averts potential misuse, venerating both data and its associated user rights.

### xii. Request for Context

Embedding the "Request for Context" ethos in GenAI platforms demands an unobstructed dialogue between the user and the system, reminiscent of the interdependencies in sociotechnical systems.[303] A GenAI news aggregator, for instance, could unambiguously expound its reliance on a user's reading trajectory to curate tailored articles, fostering enlightened user interaction. Symmetrically, platforms might elicit context, such as a GenAI chatbot soliciting the essence of a user's query, to better align its output. This reciprocity not only vests users with informed autonomy but hones AI outputs using the guiding light of user-context.

### xiii. Protection of Anonymity

Anonymity is pivotal for bolstering user trust in GenAI platforms.[304] For instance, a GenAI feedback tool might strip all user reviews of

[300] Article 17 of the General Data Protection Regulation. (n.d.). Retrieved from <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32016R0679>.

[301] Kantarcioglu, M., & Dasgupta, K. (2019). The right to erasure in the age of artificial intelligence: Challenges and opportunities. *IEEE Security & Privacy*, *17*(4), 78-85.

[302] Greenberg, M. (2023, May 2). ChatGPT users can now ask OpenAI to delete their data. TechCrunch. Retrieved from <https://techcrunch.com/2023/05/02/chatgpt-delete-data/?guccounter=1>.

[303] Mulligan, D. K., & King, J. L. (2011). Bridging the gap between privacy and design. *University of Pennsylvania Law Review*, 159(5), 1087-1174.

[304] Barocas, S., & Nissenbaum, H. (2014). Privacy and discrimination: Why anonymization does not work. *The John D. and Catherine T. MacArthur Foundation Series on Digital Media and Learning*. New York, NY: NYU Press.

personally identifiable information, letting users offer genuine insights without jeopardising their identity. If a GenAI health platform stores data, it should not only adopt advanced anonymisation techniques but also consider differential privacy — a method that adds statistical noise to datasets, ensuring individual data points remain indistinguishable.[305] This approach not only lessens the ramifications of potential data breaches but also underscores the platform's dedication to shielding user identity. Crucially, even when data is anonymised, platforms must exercise meticulous responsibility to avert any re-identification risks.[306]

### xiv. Best interest of the child

Given the impact of technology on children, technology laws globally dedicate provisions on protecting the interest of this class. Ensuring the best interest of the child is not a pure-play privacy principle, yet it finds its place in our framework owing to how the GenAI technology interacts with the society with a clear impact on this class. GenAI platforms too must implement robust filtering mechanisms to meticulously exclude unsuitable content from children's datasets. It is essential to augment these safeguards with parental control tools, offering guardians oversight over AI interactions with minors.[307] Collaborating with child psychology and tech ethics experts can further ensure ethically sound AI training, respecting the nuances of children's data. By setting clear boundaries on data personalisation, especially for users identified as children, GenAI platforms can mitigate undue influence and thereby bolster trust, positioning themselves as responsible custodians in the AI landscape.[308] According to the proposed EU AI Act, GenAI platforms have a moral and legal obligation to abstain from generating malicious or inappropriate content, including hate speech or child exploitation material.[309] Measures to forestall the creation of deep fakes— manipulated visual

[305] Choudhary, S., & Aggarwal, C. C. (2017). Differential privacy in healthcare: A survey. IEEE Transactions on Knowledge and Data Engineering, 29(9), 2189-2210.

[306] Gupta, A., & Kagal, L. (2020). Preventing deanonymization in GenAI platforms: A survey of techniques and challenges. *ACM Computing Surveys*, 53(2), 1-35.

[307] Crawford, K., & Schultz, J. (2014). Big data and the child protection imperative. *New Media & Society*, 16(1), 196-214.

[308] Crawford, K., & Schultz, J. (2014). Big data and the child protection imperative. *New Media & Society*, 16(1), 196-214.

[309] Chapter 3 European Union. (2022). Regulation (EU) 2022/765 of the European Parliament and of the Council on artificial intelligence (Artificial Intelligence Act). Official Journal of the European Union.

or auditory content portraying fictitious actions or statements— are also paramount.

### xv. Accountability and Oversight

Accountability is non-negotiable for GenAI platform developers. The proposed EU AI Act mandates that these high-risk AI systems must possess the capability to elucidate the platform's functionality and its decision-making rationale, ensuring that they can be held liable for any inadvertent damages instigated by the system.[310] While these platforms can generate unpredictable outputs, or hallucinate,[311] established accountability systems offer mechanisms to address anomalies or adverse outcomes.

To foster accountability, GenAI platforms must also prioritise "explainability", providing transparent insights into their decision-making processes.[312] Such transparency fosters trust among users, be they individuals, businesses, or regulators, and facilitates the platform's continuous improvement. Moreover, elucidating AI outputs empowers professionals across sectors, enhancing their tool's efficacy. According to the proposed EU AI Act, GenAI platforms are obligated to maintain transparency, empowering users with a clear understanding of its operational mechanisms and decision-making processes.[313] This entails offering clarity on the training data, the incorporated algorithms, and the platform's inherent risks and benefits. As AI's societal role magnifies, its explainability remains crucial for fostering informed public discourse.

It is imperative that AI's inherent autonomy does not equate to irresponsibility; there must always be a human or organisational body accountable for its actions. Incorporating feedback loops within an accountability framework allows for constant assessment

[310] European Parliament. (2023). The European Union's Artificial Intelligence Act: A briefing. Retrieved from <https://www.europarl.europa.eu/RegData/etudes/BRIE/2023/739342/EPRS_BRI(2023)739342_EN.pdf>.

[311] OpenAI. (2023). GPT-4 Technical Report. Retrieved from <https://cdn.openai.com/papers/gpt-4.pdf>.

[312] Barredo Arrieta, A., Botta, A., Donini, P., & Ivkovic, M. (2020). Explainable AI: Concepts, taxonomies, opportunities and challenges towards responsible AI. *Artificial Intelligence*, 277, 1-35.

[313] European Commission. (2020). AI and interpretability: Policy briefing. Retrieved from <https://ec.europa.eu/futurium/en/system/files/ged/ai-and-interpretability-policy-briefing_creative_commons.pdf>.

and iterative refinement of the AI's decisions, ensuring its ongoing improvement and reliability.

### xvi. Risk Assessment

Given the complexity and the varied use-cases in which GenAI can be deployed, it is crucial to assess the potential risks emanating from its use.[314] According to the proposed EU AI Act, GenAI platform developers are required to curate and perpetually update a technical dossier that delineates the system's architecture and its conformity with the EU AI Act's criteria.[315] This compilation must be accessible not only to its users but also to regulatory bodies upon demand. Moreover, prior to their market introduction or activation, GenAI platforms are mandated to undergo a conformity assessment. This scrutiny is performed by a notified body, an independent entity accredited by the European Union for such evaluations.

## 7.6 Conclusion

This working paper sheds light on a pivotal realm in the contemporary technological landscape – the privacy dimensions of GenAI platforms. Through arduous investigation, we identified sixteen privacy principles, acting as a touchstone against which GenAI platforms ought to be measured. Nevertheless, the evolutionary nature of technology and privacy concerns necessitates the continual refinement of this framework. Engaging subject matter experts to augment and finetune the principles will be a crucial next step.

While we delineate four distinct stages of the AI lifecycle in our study, we recognise the fluidity of these stages in practice. As such, based on cogent feedback from stakeholders, we are inclined to delve deeper, either subdividing these stages for greater clarity or potentially introducing an additional stage to the lifecycle. This

[314] European Commission. (n.d.). Regulatory framework for AI. *Digital-Strategy.ec.europa.eu*. Retrieved from <https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai>.

See also: Challen, K., & Jones, N. (2023). Risk management in the artificial intelligence act. *European Journal of Risk Regulation*, 4(1), 25-42. Retrieved from <https://www.cambridge.org/core/journals/european-journal-of-risk-regulation/article/risk-management-in-the-artificial-intelligence-act/2E4D5707E65EFB3251A76E288BA74068>.

[315] Chapter 3 European Union. (2022). Regulation (EU) 2022/765 of the European Parliament and of the Council on artificial intelligence (Artificial Intelligence Act). Official Journal of the European Union.

iterative approach underscores our commitment to ensuring that the framework remains both robust and adaptive.

Our ambitions do not halt at framework development. In subsequent phases of our research, we envisage applying the refined framework to a representative selection of GenAI platforms. By doing so, we aim to ascertain their performance against our meticulously curated criteria. Through this holistic exploration, we aspire not only to establish a benchmark in the GenAI privacy domain but also to fill the discernible gaps in research, thereby contributing substantively to the broader discourse on technology and privacy.

## 7.7 ANNEXURE 1

| S. No. | Privacy Principle | Stage | | | |
|---|---|---|---|---|---|
| | | Design | Development | Deployment | Audit |
| 1 | PbD1: Proactive not Reactive; Preventative not Remedial | Y | Y | Y | N |
| 2 | PbD2: Privacy as the Default Setting | Y | N | Y | Y |
| 3 | PbD3: Privacy Embedded into Design | Y | Y | N | N |
| 4 | PbD4: Full Functionality — Positive-Sum, not Zero-Sum | Y | Y | Y | Y |
| 5 | PbD5: End-to-End Security — Full Lifecycle Protection | Y | Y | Y | Y |
| 6 | PbD6: Visibility and Transparency — Keep it Open | Y | Y | Y | Y |
| 7 | PbD7: Respect for User Privacy — Keep it User-Centric | Y | Y | Y | N |
| 8 | Notice | Y | N | Y | Y |
| 9 | Data Minimisation | Y | N | Y | N |
| 10 | Purpose Limitation | Y | Y | Y | Y |
| 11 | Right to Erasure | Y | Y | Y | Y |
| 12 | Request for Context | Y | Y | N | N |
| 13 | Anonymity | Y | Y | Y | Y |
| 14 | Best interest of the child | Y | Y | Y | Y |
| 15 | Accountability and Oversight | Y | Y | N | Y |
| 16 | Risk Assessment | Y | N | Y | Y |

# 8 Towards Trustworthy AI: Guidelines for Operationalisation and Responsible Adoption

**Ms Rama Vedashree, Former CEO at Data Security Council of India;**
**Ms Jameela Sahiba, Senior Program Manager, The Dialogue;**
**Ms Bhoomika Agarwal, Research Associate, The Dialogue.**

## Abstract

Amid the rapid evolution of artificial intelligence (AI), the need for a trust-based governance framework has gained prominence. While AI promises substantial benefits, its responsible integration demands meticulous attention due to its intricate, often inscrutable nature. In contrast to traditional technologies, AI's dynamic behaviour and potential biases raise concerns regarding ethics, fairness, and unintended consequences. This paper advocates for a principled governance model to ensure responsible AI adoption. In the context of the evolving AI landscape, the paper serves the purpose of converting the widely accepted principles of trustworthy AI into tangible, actionable steps designed for both AI developers and AI users. Further, the paper provides a comprehensive approach that addresses both the technical and non-technical dimensions. The technical layer of the strategies is dedicated to crafting practical and deployable solutions for integrating trustworthy AI into intricate systems. This involves designing mechanisms that ensure transparency, fairness, and accountability within AI's intricate workings. In parallel, the non-technical layer delves into pioneering incentive strategies that cultivate a climate of conscientious AI adoption. This layer actively contributes to building a sustainable framework for AI utilisation by encouraging ethical practices and responsible decision-making.

## 8.1 Introduction

In recent years, the rapid advancement of artificial intelligence (AI) has sparked a global conversation about the ethical implications of this transformative technology. AI technologies have the potential to bring about significant benefits, but their responsible and ethical deployment requires careful consideration. Unlike traditional technologies, AI operates in a dynamic and often opaque manner, making it challenging to understand and predict its behavior. This raises concerns about biases, discrimination, and unintended consequences that can have far-reaching societal impacts.

Discussions surrounding regulating this fast-paced technology often revolve around the delicate balance between mitigating potential risks and promoting innovation and adoption. Central to achieving this equilibrium is establishing a principle-based congruence on a global level, which is essential in building the required public trust for harnessing the full benefits of this technology. These principles will serve as guiding values, shaping the development and implementation of regulatory frameworks pertaining to AI. To accomplish the same, in the first part of the paper, we map and analyse trustworthy AI principles by conducting a comprehensive landscape study of regulatory frameworks from around the world.

In the second part of the paper, we aim to develop an operational strategy to translate the identified principles into action points. For the purposes of this paper, we will focus on mapping the operationalization process to two key stakeholders: AI developers and AI users, both at technical and non-technical levels. The technical aspect of operationalization will focus on providing technical solutions to adopt trustworthy AI principles at the level of AI developers and users. This will include developing action points that are technically feasible and can be implemented within AI systems.

The aim is to offer practical approaches and solutions that address the challenges of implementing trustworthy AI. The non-technical aspect of operationalization will involve strategies to incentivize AI developers and users to adopt trustworthy practices. This will help recognize that while technical solutions are essential, it is equally important to motivate and encourage stakeholders who may not possess specialized technical knowledge. The paper will explore various incentive mechanisms and strategies that can effectively encourage AI developers and users to embrace and implement trustworthy AI principles.

## 8.2 Typology of Principles

This section embarks on a landscape study of the various ethical AI frameworks that have emerged across the globe. The landscape study will help identify a typology of trustworthy AI principles specifically tailored to the objectives of this paper. With the abundance of ethical AI frameworks, it becomes crucial to identify the key elements that contribute to the credibility and effectiveness of these principles. By studying the landscape of AI ethical frameworks, we can identify the core principles that consistently emerge across multiple frameworks, thereby enabling the formulation of a typology of trustworthy AI principles.

## 8.3 Landscape study of several frameworks

### 8.3.1 OECD AI Principles

OECD AI Principles are a set of internationally agreed principles that seek to promote human-centric AI. The document is divided into two parts: first, it delineates five key principles that all AI actors are encouraged to adopt for responsible stewardship of trustworthy AI. These principles include: a) Inclusive growth, sustainable development and well-being, b) Human-centred values and fairness, c) Transparency and explainability d) Robustness, security and safety, and e) Accountability. The document stresses the complementary nature of these principles. The second part of the legal instrument lays down recommendations for countries to help them implement the above-mentioned principles. The recommendations range from facilitating investment in R&D for fostering innovation in trustworthy AI to framing enabling policies and increased cooperation at international forums.

### 8.3.2 G20 AI Principles

Drawing reference from OECD principles, the G20 also adopted identical principles for responsible stewardship of trustworthy AI in June 2019, so as to promote and implement: (a) inclusive growth, sustainable development and well-being, (b) human-centred values and fairness, (c) transparency and explainability, (d) robustness, security and safety, and (e) accountability. The aim is to foster beneficial outcomes, including augmenting human capabilities, reducing inequalities, and protecting the environment. The principles of transparency and responsible disclosure enable informed decision-making, while robustness, security, and safety mitigate risks. These principles emphasize traceability, risk management, and accountability, addressing concerns such as privacy, digital security, safety, and bias.

### 8.3.3 EU Ethics Guidelines for Trustworthy AI

The European Commission constituted a High-Level Expert Group on Artificial Intelligence to develop guidelines for the promotion of trustworthy AI. The Guidelines identify three components of trustworthy AI: lawful, ethical and robust. Using fundamental rights as the basis for developing trustworthy AI, the guidelines devise four ethical principles that should be adhered to during the development, deployment and usage of AI : (i) Respect for human autonomy, (ii) Prevention of harm, (iii) Fairness (iv) Explicability (transparency, openness, explainability). Building on these principles, seven requirements are delineated that can be met through technical and non-technical methods. These include: a)Human agency and oversight, b) Technical robustness and safety, c) Privacy and data governance, d) Transparency, e) Diversity, non-discrimination and fairness, f) Societal and environmental well-being , g) Accountability. The guidelines further provide an assessment list for the actors to ensure that the AI complies with these principles. The guidelines acknowledge the possibility of potential conflicts between principles and emphasises the need for determining trade-offs based on evidence and reason.

### 8.3.4 UNESCO Ethics of Artificial Intelligence

The UNESCO Ethics of Artificial Intelligence framework encompasses several key principles that aim to guide the responsible development and deployment of AI technologies. These principles address various aspects of AI systems, focusing on ensuring ethical practices and upholding human rights and fundamental freedoms. The key principles include: (a) Proportionality and doing no harm; (b) Safety and Security; (c) Fairness and Non-discrimination; (d) Sustainability; (e) Right to Privacy and Data Protection; (f) Human oversight and determination; (g) Transparency and explainability; (h) Responsibility and accountability; (i) Awareness and literacy; and (j) Multi-stakeholder and adaptive governance and collaboration. The objectives of the UNESCO Ethics of Artificial Intelligence are to establish a universal framework of values, principles, and actions that guide states in formulating AI-related legislation and policies in accordance with international law.

## 8.4 Mapping Trustworthy AI Principles

Through an extensive analysis of various ethical AI frameworks worldwide, it has become evident that certain principles play a pivotal role in ensuring the development and deployment of trustworthy AI technology. It is also important to acknowledge that not all of them are centered on promoting ethical or trustworthy AI. Certain frameworks outlined above place a greater emphasis on AI regulation and governance, underscoring the importance of adhering to legal and operational standards, in contrast to those that are centred on establishing ethical guidelines to promote trustworthy and responsible AI. However, within this diverse array of frameworks, our synthesis has uncovered a set of core principles that consistently surface across different contexts. These principles underscore their fundamental importance regardless of the framework's primary focus. They often revolve around concepts such as transparency, accountability, fairness, and the protection of individual rights and privacy. In the next section, we go into further detail into how these principles can be operationalised.

## 8.5 Operationalisation of Trustworthy AI Principles

As various stakeholders strive to embrace AI's potential, there arises a pressing need to develop a comprehensive operational strategy that translates identified principles into actionable steps. Our methodology, drawing from an array of ethical guidelines and best practices, goes beyond mere theoretical discussions. We delve into practical implementation, at both technical and non-technical levels. The operationalization process, as explained in this paper, focusses on two key participants: AI developers and AI users. By addressing their specific needs and responsibilities, we aim to foster a culture of trustworthy AI adoption, accountability, and transparency.

Through this approach, we aspire to demonstrate the universal relevance of our strategy and encourage its adoption across sectors, ultimately fostering a responsible and ethical AI ecosystem for the betterment of society as a whole. We explore the technical and non-technical strategies to operationalize the identified trustworthy AI principles. At the technical level, it will delve into specific AI development techniques and practices that align with each principle. On the non-technical side, the section will focus on policy and governance approaches to incentivize AI developers and users to adhere to trustworthy practices. The aim is to present a comprehensive and balanced perspective on operationalizing trustworthy AI principles from both technical and non-technical angles.

## 8.6 Transparency and Explainability

Transparency in AI allows modelers, developers, and technical auditors to gain a comprehensive understanding of the AI system's intricacies, including training, evaluation, decision boundaries, input processing, and the reasoning behind specific predictions.[316] Building upon transparency, Explainable AI (XAI)[317] goes a step further, providing extensive explanations to users and customers, elucidating the system's functioning and the logic behind specific recommendations. The quest for explainability stems from the need to demystify the black-box nature of AI algorithms and provide meaningful insights to stakeholders.[318] A clear understanding of the decision-making processes enables researchers to validate AI systems rigorously and identify potential biases or errors that may arise during model operation.

[316] Building Transparency into AI Projects. (2022, June 20). Harvard Business Review. Retrieved August 25, 2023, from <https://hbr.org/2022/06/building-transparency-into-ai-projects>.

[317] What is explainable AI? (n.d.). IBM. Retrieved August 25, 2023, from <https://www.ibm.com/watson/explainable-ai>.

[318] Vorras, A., & Mitrou, L. (n.d.). Unboxing the Black Box of Artificial Intelligence: Algorithmic Transparency and/or a Right to Functional Explainability. EU Internet Law in the Digital Single Market, 2021. <https://link.springer.com/chapter/10.1007/978-3-030-69583-5_10/>.

### 8.6.1 Technical Level

At the technical level, operationalizing transparency and explainability in AI systems involves adopting model interpretability techniques to shed light on the decision-making processes of AI models. Techniques like LIME (Local Interpretable Model-agnostic Explanations)[319] and SHAP (Shapley Additive Explanations)[320] are valuable tools that AI developers can leverage. LIME generates locally interpretable models to explain individual predictions, allowing developers to gain insights into the factors that contribute to specific outcomes. On the other hand, SHAP provides a game-theoretic approach to explain the output of any machine learning model, attributing the contribution of each input feature to the final prediction. By utilizing these methods, AI developers can enhance their understanding of the model's inner workings, making it easier to identify potential biases, errors, or sources of unethical behavior.

In addition to model interpretability techniques, user-friendly dashboards play a crucial role in enhancing transparency for end-users.[321] These dashboards present AI outputs in a clear and understandable manner, allowing users to comprehend how the AI system arrives at specific decisions or recommendations. By providing comprehensive explanations, users can gain trust and confidence in the AI technology, making it more user-friendly and accessible. The transparency achieved through such dashboards not only fosters user trust but also empowers users to make informed decisions based on AI-driven insights.

Another important aspect of operationalizing transparency is ensuring data lineage[322] and maintaining detailed documentation of AI model development. Data lineage enables AI developers and technical auditors to trace the origin and transformation of data

[319] Huang, A., Li, J., & Shankar, N. (2020, August 31). 6 – Interpretability – Machine Learning Blog | ML@CMU | Carnegie Mellon University. ML@CMU Blog. Retrieved August 25, 2023, from <https://blog.ml.cmu.edu/2020/08/31/6-interpretability/>.

[320] Huang, A., Li, J., & Shankar, N. (2020, August 31). 6 – Interpretability – Machine Learning Blog | ML@CMU | Carnegie Mellon University. ML@CMU Blog. Retrieved August 25, 2023, from <https://blog.ml.cmu.edu/2020/08/31/6-interpretability/>.

[321] InterpretML: A toolkit for understanding machine learning models*. (2020, May 18). Microsoft. Retrieved August 25, 2023, from <https://www.microsoft.com/en-us/research/uploads/prod/2020/05/InterpretML-Whitepaper.pdf>.

[322] What is Data Lineage? (n.d.). Informatica. Retrieved August 25, 2023, from <https://www.informatica.com/resources/articles/what-is-data-lineage.html>.

throughout the AI system's life cycle. This helps in understanding how data inputs are processed and used within the model, leading to greater clarity on how the AI system generates predictions. Detailed documentation of AI model development provides crucial information about the model's architecture, training data, hyperparameters, and evaluation metrics.[323] This documentation promotes transparency by enabling other researchers and auditors to validate the AI system's performance and scrutinize its decision-making processes.

### 8.6.2 Non-technical Level

To promote transparency and encourage AI developers to prioritize explainability, regulators can implement several strategies. One effective approach is to mandate transparency reporting requirements for AI systems deployed in critical sectors such as finance and healthcare.[324] These reporting requirements would compel AI developers to provide detailed information about their AI models, including the data used for training, the decision-making processes, and any potential biases or limitations of the system. By making this information publicly available, stakeholders and users can gain insight into the inner workings of the AI system, which fosters trust and accountability.

Furthermore, providing incentives to AI developers who adhere to transparency standards can be a powerful motivator. Regulators can offer certification or accreditation programs for AI systems that meet specific transparency criteria.[325] AI developers who attain these certifications can showcase their commitment to transparency and differentiate their products in the market. This can create a competitive advantage for transparent AI systems, incentivizing developers to prioritize explainability in their AI models.

[323] Konigstorfer, F., & Thalmann, S. (n.d.). AI Documentation: A path to accountability. Journal of Responsible Technology, 11(2022)100043. <https://www.sciencedirect.com/science/article/pii/S2666659622000208/>.

[324] Microsoft Responsible AI Standard, v2: General Requirements. (n.d.). The Official Microsoft Blog. Retrieved August 25, 2023, from <https://blogs.microsoft.com/wp-content/uploads/prod/sites/5/2022/06/Microsoft-Responsible-AI-Standard-v2-General-Requirements-3.pdf>.

[325] IEEE SA. (2023, July 31). The Ethics Certification Program for Autonomous and Intelligent Systems (ECPAIS) – IEEE Standards Association. Retrieved from <https://standards.ieee.org/industry-connections/ecpais/>.

In addition to regulatory measures and incentives, awareness campaigns and educational initiatives targeted at users can play a significant role in fostering a culture of demand for transparent AI systems.[326] Many users may not fully understand the implications of AI technology and the importance of transparency. Educating users about the benefits of transparent AI systems and the potential risks of opaque models can empower them to demand more accountability from AI developers. This increased demand for transparency can create a market-driven push for AI developers to be more transparent and user-centric in their approach.

## 8.7 Accountability

Accountability is a critical principle that underpins the entire lifecycle of an AI system.[327] It demands that all stakeholders involved in the development and deployment of AI systems take responsibility for ensuring that the technology aligns with human values. This accountability is achieved through careful product design, reliable technical architecture, a thorough assessment of potential impacts, and the transparent disclosure of information related to these aspects. Transparency plays a fundamental role in facilitating the accountability of an AI system by providing the means to understand and justify its decisions and actions. Derived from accountability, the concept of auditability also comes into play, requiring that the justification of an AI system be subject to review, assessment, and auditing.[328]

### 8.7.1 Technical Level

At the technical level, ensuring accountability and auditability in AI systems is crucial for instilling trust and confidence among users and stakeholders. By holding developers and operators accountable for their design and implementation choices, the risk of biased or unethical AI outcomes can be mitigated. To operationalize

[326] Endsley, M. R. (2023). Supporting Human-AI Teams:Transparency, explainability, and situation awareness. Computers in Human Behavior, 140, 107574. <https://doi.org/10.1016/j.chb.2022.107574>.

[327] Novelli, C., Taddeo, M., & Floridi, L. (2023). Accountability in artificial intelligence: what it is and how it works. AI & Society. <https://doi.org/10.1007/s00146-023-01635-y>.

[328] Williams, R., Cloete, R., Cobbe, J., Cottrill, C., Edwards, P., Markovic, M., Pang, W. (2022). From transparency to accountability of intelligent systems: Moving beyond aspirations. Data & Policy, 4. <https://doi.org/10.1017/dap.2021.37>.

accountability and auditability, organizations should establish clear governance frameworks and mechanisms for oversight. Establishing clear accountability frameworks is of paramount importance in the development of AI systems. AI developers must take proactive steps to define roles and responsibilities for each stakeholder involved in the AI development process. This ensures that everyone understands their obligations and is accountable for their respective contributions to the AI system. By delineating responsibilities, developers can identify key decision-makers, data handlers, and model architects, making it easier to attribute outcomes and actions to specific individuals or teams.[329]

One effective way to promote accountability is by implementing robust audit trails and logs throughout the AI system's life cycle. These audit trails record and track every action, decision, and modification made within the AI system. By maintaining detailed records, developers can trace the decision-making process back to individual contributors, thereby enhancing transparency and facilitating accountability.[330] Audit trails also serve as a valuable tool for identifying potential issues, biases, or errors in the AI system, enabling developers to take corrective actions promptly.

Another way is through Algorithmic auditing,[331] which is a recognized approach to ensure accountability and assess the impact of an AI system on various dimensions of human values. This auditing process involves evaluating the AI system's algorithms, data inputs, and decision-making processes to identify potential biases, ethical considerations, and compliance with regulations and ethical guidelines.

### 8.7.2 Non-technical Level

At the non-technical level, encouraging AI developers and users to adopt ethical guidelines and industry best practices is instrumental

[329] Williams, R., Cloete, R., Cobbe, J., Cottrill, C., Edwards, P., Markovic, M., Pang, W. (2022c). From transparency to accountability of intelligent systems: Moving beyond aspirations. Data & Policy, 4. <https://doi.org/10.1017/dap.2021.37>.

[330] Ali, S., Abuhmed, T., El-Sappagh, S., Muhammad, K., Alonso-Moral, J. M., Confalonieri, R., . . Herrera, F. (2023). Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence. Information Fusion, 99, 101805. <https://doi.org/10.1016/j.inffus.2023.101805>.

[331] <https://arxiv.org/pdf/2001.00973.pdf>.

in fostering accountability.[332] These guidelines serve as a compass, guiding developers and users to make decisions that align with ethical principles and societal values. By adhering to these principles, developers and users can ensure that AI systems are developed and utilized responsibly, minimizing the risk of biased or harmful outcomes.

Regulatory mechanisms play a crucial role in holding organizations accountable for the consequences of AI decisions.[333] Implementing regulations that define the responsibilities and liabilities of organizations in deploying AI technologies reinforces the importance of ethical considerations and encourages compliance. Such mechanisms act as powerful incentives for organizations to prioritize transparency, fairness, and safety in their AI systems, as they become legally bound to be answerable for any adverse impact resulting from AI actions.

Further, to foster a culture of accountability and responsible AI development, continuous training and education are also essential.[334] Developers and operators need to stay updated on the latest developments in AI ethics and best practices to ensure that they make informed decisions throughout the AI system's life cycle. Providing ongoing training helps to instill a sense of responsibility and ownership, emphasizing the significance of adhering to ethical guidelines and industry standards. In addition, organizations must embrace industry standards and regulations specific to AI development and usage.[335] Compliance with these standards ensures that AI systems undergo rigorous scrutiny and assessment to meet predefined criteria for fairness, explainability, and safety.

## 8.8 Fairness and Non-discrimination

The utilization of AI systems in critical areas like health, financial risk assessment, recruitment and face identification has brought

[332] Ryan, M., & Stahl, B. C. (2020). Artificial intelligence ethics guidelines for developers and users: clarifying their content and normative implications. Journal of Information, Communication and Ethics in Society, 19(1), 61–86. <https://doi.org/10.1108/jices-12-2019-0138>.

[333] Sanford, S. (2021, August 30). How to Build Accountability into Your AI. Retrieved from <https://hbr.org/2021/08/how-to-build-accountability-into-your-ai>.

[334] Responsible AI | AI Ethics & Governance. (n.d.). Retrieved from <https://www.accenture.com/in-en/services/applied-intelligence/ai-ethics-governance>.

[335] Theoto, T., Küspert, S., Hefter, K., Mills, S., Bickford, J. K., Malik, P., Roselund, T. (2023). Responsible AI for an era of tighter regulations. BCG Global. Retrieved from <https://www.bcg.com/publications/2022/acting-responsibly-in-tight-ai-regulation-era>.

attention to the potential consequences of systematic unfairness and discrimination in AI decisions.[336] These biases can lead to negative social impacts, as disadvantaged groups may face systematic disadvantages. Such biases not only erode trust in AI but also hinder the technology's overall potential to benefit society. Consequently, practitioners must prioritize the fairness of AI systems to avoid perpetuating or exacerbating social bias. Two key factors that contribute to bias are group identity (sensitive variables) and the system's response (prediction).

### 8.8.1 Technical Level

At the technical level, operationalizing fairness and non-discrimination in AI systems involves the application of various techniques aimed at mitigating biases and promoting equitable outcomes. Debiasing algorithms is one such approach, that seeks to identify and address biases present in the data or the model itself.[337] These algorithms can adjust the training data or modify the model's parameters to reduce bias and ensure fair treatment across different groups. Another technique is adversarial training, where AI systems are exposed to adversarial scenarios designed to simulate real-world challenges related to bias. By subjecting the AI model to these scenarios, the system learns to be resistant to bias and makes fair predictions even in challenging circumstances.

Fairness-aware learning is another method that explicitly incorporates fairness constraints during the training process.[338] This approach involves considering fairness as an integral part of the AI model's objective function, ensuring that fairness is optimized alongside accuracy and other performance metrics. By incorporating fairness and non-discrimination as a core criterion, developers can design AI systems that inherently prioritize fair outcomes.

[336] Hunkenschroer, A. L., & Kriebitz, A. (2022). Is AI recruiting (un)ethical? A human rights perspective on the use of AI for hiring. AI And Ethics, 3(1), 199–213. <https://doi.org/10.1007/s43681-022-00166-4>.

[337] Xu, J. (2021, December 10). Algorithmic Solutions to Algorithmic Bias: A Technical Guide. Medium. Retrieved from <https://towardsdatascience.com>.

[338] Jin, D., Wang, L., He, Z., Zheng, Y., Ding, W., Xia, F., & Pan, S. (2023). A survey on fairness-aware recommender systems. Information Fusion, 100, 101906. <https://doi.org/10.1016/j.inffus.2023.101906>.

### 8.8.2 Non-technical Level

Addressing fairness and non-discrimination in AI requires a holistic approach that extends beyond technical solutions. Collaboration among experts from various disciplines is crucial for understanding the broader societal implications of AI decisions. Ethicists, legal experts, sociologists, and others can contribute their expertise to define fairness criteria that aligns with societal values and norms.[339] This interdisciplinary approach ensures that AI systems are designed and deployed in a manner that considers ethical and societal considerations.

Further, implementing diversity and inclusion policies within AI development teams is a critical step towards achieving fairness in AI. In addition, raising awareness about the impact of biased AI decisions is vital in promoting fairness and non-discrimination. Public campaigns and educational initiatives can help inform the public about the potential consequences of biased AI systems, generating social pressure for developers and organizations to prioritize fairness and non-discrimination.[340] Increased awareness can also empower individuals to demand fair and transparent AI solutions.

## 8.9 Reliability and Safety/Robustness

Reliability and safety/robustness are fundamental principles in ensuring the trustworthy operation of AI systems.[341] Reliability refers to the ability of an AI algorithm or system to consistently perform accurately under varying conditions and inputs. A reliable AI system should produce consistent and dependable results, instilling confidence in its users and stakeholders. Banking on reliability, robustness goes further ahead and encompasses the ability of an AI system to handle unexpected situations, errors, or erroneous

[339] Mantelero, A. (2022). The social and ethical component in AI systems design and management. In Information technology & law series (pp. 93–137). <https://doi.org/10.1007/978-94-6265-531-7_3>.

[340] Tackling bias in artificial intelligence (and in humans). (2019, June 6). Retrieved from <https://www.mckinsey.com/featured-insights/artificial-intelligence/tackling-bias-in-artificial-intelligence-and-in-humans>.

[341] Msteller-Ai. (2023, July 28). Responsible and trusted AI – Cloud Adoption Framework. Retrieved from <https://learn.microsoft.com/en-us/azure/cloud-adoption-framework/innovate/best-practices/trusted-ai#:~:text=Reliability%20and%20safety,-For%20AI%20systems&text=An%20organization%20should%20establish%20rigorous,performance%20can%20degrade%20over%20time>.

inputs gracefully.[342] A robust AI system can adapt to dynamic and diverse environments and still produce reliable results. It should be resilient to variations in data, changes in input distributions, or the presence of outliers.

### 8.9.1 Technical Level

To achieve reliability/safety/robustness in AI systems, developers employ a variety of techniques that strengthen the system's ability to perform reliably and accurately in diverse and challenging situations. One such technique is data augmentation, where the training data is enriched with various transformations and perturbations. By exposing the model to a broader range of data distributions, data augmentation helps the AI system generalize better and handle unseen data more effectively.

Adversarial training is another powerful approach used to enhance robustness. Adversarial attacks involve deliberately introducing small perturbations to input data that can cause the AI model to produce incorrect or misleading outputs.[343] Through adversarial training, the AI system is trained to recognize and defend against these adversarial inputs, making it more resilient to potential attacks.

Further, uncertainty estimation is a crucial aspect of achieving robustness in AI. AI systems must be able to recognize situations where their predictions may be uncertain or less reliable.[344] Uncertainty estimation techniques help quantify and communicate the confidence levels of the AI model's predictions, enabling appropriate caution or human intervention in critical scenarios.

At the technical level, stress testing and scenario analyses also play a pivotal role in evaluating the robustness of AI systems. Stress testing involves subjecting the AI model to extreme or challenging conditions to assess its performance under adverse circumstances. Scenario analyses, on the other hand, explore how the AI system responds

[342] Singh, R. (2020, November 2). Trustworthy AI. Retrieved from <https://arxiv.org/abs/2011.02272>.

[343] Goyal, S., Doddapaneni, S., Khapra, M. M., & Ravindran, B. (2023). A survey of Adversarial Defenses and Robustness in NLP. ACM Computing Surveys, 55(14s), 1–39. <https://doi.org/10.1145/3593042>.

[344] Abdar, M., Pourpanah, F., Hussain, S., Rezazadegan, D., Liu, L., Ghavamzadeh, M., . . Nahavandi, S. (2021). A review of uncertainty quantification in deep learning: Techniques, applications and challenges. Information Fusion, 76, 243–297. <https://doi.org/10.1016/j.inffus.2021.05.008>.

to specific hypothetical situations, enabling developers to identify potential weaknesses and areas for improvement. Implementing error monitoring mechanisms is vital to detect and address issues promptly. By continuously monitoring AI system performance, developers can identify anomalies and errors early on, allowing for timely intervention and rectification.

### 8.9.2 Non-technical Level

On the non-technical side, providing financial incentives for organizations that prioritize and maintain robust AI systems can be a powerful motivator.[345] Governments and regulatory bodies can offer grants, tax benefits, or other financial rewards to organizations that demonstrate a commitment to reliability and safety in their AI deployments. These incentives can encourage businesses to invest in robustness and allocate resources to continuously monitor and improve their AI systems' performance.

Transparency and clear communication between regulatory bodies and AI developers is also crucial.[346] Regular consultations and open dialogues can facilitate a better understanding of each other's perspectives and concerns. This enables regulators to gain deeper insights into AI technologies' complexities, allowing them to design more effective policies and standards. Similarly, AI developers can gain clarity on regulatory expectations, which helps them align their practices with safety and reliability goals.

## 8.10 Privacy and Data Protection

A commitment to privacy protection is essential because it not only respects individuals' rights to privacy but also plays a crucial role in determining the overall trustworthiness of an AI system.[347] When users entrust their data to AI systems, they expect that their personal information will be handled with utmost care and

[345] <https://plus.google.com/+UNESCO. (n.d.). Recommendation on the ethics of artificial intelligence. Retrieved from <https://en.unesco.org/about-us/legal-affairs/recommendation-ethics-artificial-intelligence>.

[346] Lawton, G. (2023). AI transparency: What is it and why do we need it? *CIO*. Retrieved from <https://www.techtarget.com/searchcio/tip/AI-transparency-What-is-it-and-why-do-we-need-it>.

[347] Reinhardt, K. (2022). Trust and trustworthiness in AI ethics. AI And Ethics. <https://doi.org/10.1007/s43681-022-00200-5>.

confidentiality. Any compromise in data privacy can lead to breaches of trust and undermine the credibility of the AI system and the organizations behind it.

### 8.10.1 Technical Level

To ensure robust privacy protection in AI systems, developers and organizations must implement a range of data privacy measures.[348] One of the key steps is adopting encryption and access controls to safeguard data from unauthorized access. Encryption involves encoding the data in a way that can only be decrypted with a specific key, ensuring that even if unauthorized individuals gain access to the data, they cannot decipher its contents. Access controls, on the other hand, limit the users who can access certain data, reducing the risk of data breaches.

Data anonymization techniques are equally critical in preserving privacy. By anonymizing data, personally identifiable information (PII) is removed or transformed in a way that prevents direct linkage to specific individuals.[349] This ensures that even if data is accessed or shared, it cannot be traced back to individuals, providing an added layer of protection.

Further, regular data audits and vulnerability assessments are vital in maintaining privacy protection. Data audits involve comprehensive reviews of data handling practices, identifying potential weak points in data management, and ensuring compliance with privacy policies.[350] Vulnerability assessments help in proactively identifying potential security loopholes and vulnerabilities in AI systems, allowing developers to address them promptly before they are exploited.

Furthermore, compliance with prevalent data protection laws and regulations would be paramount. Beyond mere legal compliance,

[348] Van Rijmenam Csp, M. (2023). Privacy in the age of AI: Risks, challenges and solutions. Dr Mark Van Rijmenam, CSP | Strategic Futurist Speaker. Retrieved from <https://www.thedigitalspeaker.com/privacy-age-ai-risks-challenges-solutions/#:~:text=Organisations%20that%20use%20AI%20must,whose%20data%20has%20been%20compromised>.

[349] What is Data Anonymization | Techniques, Pros, Cons, and Use Cases. (n.d.). Retrieved from <https://www.k2view.com/what-is-data-anonymization/#:~:text=Data%20anonymization%20transforms%20personal%20or,privacy%20laws%20and%20heighten%20security>.

[350] Quach, S., Quach, S., Martin, K. D., Weaven, S., & Palmatier, R. W. (2022). Digital technologies: tensions in privacy and data. Journal of the Academy of Marketing Science, 50(6), 1299–1323. <https://doi.org/10.1007/s11747-022-00845-y>.

ethical considerations must be integrated into AI system design and deployment. Transparency plays a vital role in this regard, as developers should openly communicate their data collection and usage practices to users.[351] Providing clear and easily understandable explanations empowers users to make informed decisions about sharing their data, giving them greater control over their information.

Lastly, consent mechanisms are central to upholding privacy and user autonomy. Developers should implement clear and explicit consent processes, seeking users' permission before collecting and using their data.[352] Individuals should have the option to opt-in or opt-out of data sharing, ensuring that they have the freedom to participate in AI systems without feeling coerced or manipulated.

### 8.10.2 Non-technical Level

At the non-technical level, incentivizing AI developers and users to adopt data privacy and protection principles can be achieved through a multi-faceted approach. Building user trust is a crucial first step, as AI developers can gain trust by developing clear data privacy policies and openly communicating their data usage practices.[353] By providing transparency about how user data will be collected, stored, and used, users can understand the measures in place to safeguard their information, which encourages them to engage more confidently with AI systems.

Certification and recognition also play a significant role in incentivizing data privacy adherence. Recognizing organizations that demonstrate strong data protection practices through certifications or accreditations can act as a visible badge of trust, validating an organization's commitment to data privacy. Third-party certifications boost an organization's reputation and instill confidence in users, making them more likely to choose AI systems from certified organizations over others.[354]

[351] Morey, T. (2020, September 1). Customer Data: Designing for transparency and trust. Retrieved from <https://hbr.org/2015/05/customer-data-designing-for-transparency-and-trust>.

[352] Andreotta, A., Kirkham, N., & Rizzi, M. (2021). AI, big data, and the future of consent. AI & Society, 37(4), 1715–1728. <https://doi.org/10.1007/s00146-021-01262-5>.

[353] Lawton, G. (2021). The future of trust will be built on data transparency. CIO. Retrieved from <https://www.techtarget.com/searchcio/feature/The-future-of-trust-must-be-built-on-data-transparency>.

[354] Bias and ethical concerns in machine learning. (n.d.). Retrieved from <https://www.isaca.org/resources/isaca-journal/issues/2022/volume-4/bias-and-ethical-concerns-in-machine-learning>.

Ethical considerations also come into play, as organizations can adopt ethical frameworks that prioritize user consent, fairness, and responsible data handling.[355] Promoting these ethical values fosters a culture of responsible AI development and use, motivating stakeholders to adhere to data privacy principles and prioritize user rights.

## 8.11 Conclusion

As artificial intelligence continues its rapid evolution, it is imperative that we establish a trust-based governance framework to navigate the complex landscape it presents. While AI holds tremendous promise, its intricacies, dynamic behavior, and potential biases necessitate a vigilant and principled approach to ensure responsible integration. Our paper has advocated for the development of a comprehensive governance model that translates the widely accepted principles of trustworthy AI into actionable steps for both AI developers and users.

The framework we propose addresses both the technical and non-technical dimensions of responsible AI adoption. On the technical front, it emphasizes the creation of practical and deployable solutions that enhance transparency, fairness, and accountability within AI systems. These technical mechanisms are crucial in mitigating the risks associated with opaque AI decision-making processes and unintended consequences. Simultaneously, our paper highlights the significance of the non-technical layer, focusing on pioneering incentive strategies aimed at fostering a culture of conscientious AI adoption. By encouraging ethical practices and responsible decision-making, this aspect of our framework contributes to the establishment of a sustainable environment for AI utilization.

Operationalisation of principles at technical and non-technical levels detailed in this paper would involve coordination of various factors like regulatory landscape, geopolitics etc. This is essential for the seamless implementation of the principle-based multi stakeholder approach. Coordination will involve three levels of engagement, encompassing various stakeholders. First, in terms of domestic coordination, the countries will need to ensure harmonisations

[355] Morey, T. (2020b, September 1). Customer Data: Designing for transparency and trust. Retrieved from <https://hbr.org/2015/05/customer-data-designing-for-transparency-and-trust>.

amongst several domestic laws regulating digital space. For instance, in India, both the Digital Personal Data Protection Act 2023 and the upcoming Digital India Act (DIA) would effectively address the impact and risks of AI technologies. Harmonisation between both the laws and the respective implementing authorities would have to be ensured.

Second, in terms of international coordination, the importance of building regulatory consensus internationally will have to be underscored. Several regulatory developments are taking place worldwide, such as the initiatives undertaken by the European Union, the draft AI Bill in Brazil, etc. It would be imperative to establish universal consensus on fundamental aspects of AI to ensure a cohesive and harmonized approach. Third, alternative approaches for regulating AI by leveraging market mechanisms and promoting public-private coordination would have to be explored. The mechanisms and incentives that can encourage AI developers to prioritize consumer protection and safety as a value proposition, thereby fostering trustworthiness in AI systems would need greater focus.

In an era marked by the ever-expanding influence of AI, our proposed governance model serves as a guidepost, offering a roadmap for stakeholders to navigate the challenges and complexities of AI integration. It is our hope that this principled approach will pave the way for the responsible and ethical development and utilization of artificial intelligence, ultimately leading to a future where AI enriches our lives while upholding the values and principles that underpin a just and equitable society.

# PART 3

## Western Perspectives on AI Governance

# 9 The Blind Watcher: Accountability mechanisms in the Artificial Intelligence Act

**Nicola Palladino, Research Fellow at the Trinity College Dublin's Long Room Hub Arts and Humanities Research Institute.**

## Abstract

The paper delves into the crucial aspect of accountability in the realm of artificial intelligence (AI), focusing specifically on the European Union's proposed legislation, the Artificial Intelligence Act (AIA). After highlighting the transformative impact of AI on society and the need for robust governance mechanisms to mitigate potential misuses and risks associated with AI systems, the paper underscores the importance of building trust and public acceptance for AI, given its potential to reshape decision-making processes across various sectors. The paper investigates the concept of accountability, differentiating between internal and external accountability in the context of AI systems. It emphasizes that AI's multi-stakeholder nature necessitates a comprehensive accountability framework, encompassing developers, providers, users, and regulatory bodies. The discussion delves into the AIA's regulatory approach, which classifies AI applications based on risk and mandates compliance with distinct sets of requirements. The AIA's accountability mechanisms are analyzed in-depth, from risk categorization to conformity assessments, with a focus on high-risk applications. The paper concludes by acknowledging the significance of the AIA as a pioneering regulation in the AI governance landscape. However, it raises concerns about potential shortcomings, such as the limited application of accountability requirements and the potential for vested interests to influence evaluations.

## 9.1 Introduction

"Artificial Intelligence" is a label used as shorthand for an expanding 'family' of software (and hardware) systems capable of performing specific cognitive tasks by collecting, analyzing, and interpreting data, to make decisions and take actions with a certain degree of autonomy.[356] Unlike other technologies, AI is not only giving rise to a new policy field and means of power. It is also giving rise to a novel layer of governance embedded into socio-technical architectures, in which technical specifications affect human behavior, by allowing or denying some course of actions, influencing the way we make a decision, or playing a crucial role in the way relevant decision for individual and community are made.[357]

If not managed properly, AI systems may be subject to a series of misuses endangering the autonomy and integrity of individuals and communities, calling into question issues such as privacy, discrimination, manipulation, misinformation, and the erosion of democratic institutions and the effects on jobs and rights on the workplace.[358]

As testified by recent initiatives such as the EU Ethical Guidelines for Trustworthy AI, The IEEE Ethical Aligned Design, and the OECD Recommendations on AI,[359] or the UNESCO Recommendation on the Ethics of Artificial Intelligence,[360] in the past few years, stakeholders reached the awareness that the full potential of this technology is attainable only by building a trustworthy and human-centric framework.

This means that AI systems must be aligned with societal values and governed through accountable arrangements to avoid both misuse of AI applications capable of endangering people and underuse because of a lack of public acceptance (as occurred with nuclear power or GMOs).[361] Scholars also noted how stakeholders must cooperate to achieve regulation capable of ensuring predictability and legal certainty even if the debate remains open regarding the role and responsibilities of different actors and the proper balance

[356] Russell, S., and Norvig, P. (2003). *Artificial Intelligence: A Modern Approach*. Prentice Hall Series in Artificial Intelligence. Upper Saddle River, NJ: Prentice Hall, Pearson Education.

[357] Palladino, N. (2021). *Imbrigliare i giganti digitali nella rete del costituzionalismo ibrido. Spunti dall'approccio europeo alla governance dell'Intelligenza artificiale*. In M. Santaniello (Ed.), Comunicazionepuntodoc – lower the Top. Fausto Lupetti Editore. Isbn 978-88-6874-188-4.

[358] Renda, A. (2019) *Artificial Intelligence, Ethics Governance and Policy Challenges*. Brussels: CEPS; Boiler, G., (2018) *Artificial Intelligence: The Great Disruptor*. Washington, DC: The Aspen Institute.

[359] On this point, see also Lewis, D., et al. (2020). *Global Challenges in the Standardization of Ethics for Trustworthy AI*, *Journal of ICT*, 8(2),123–150.

[360] <https://unesdoc.unesco.org/ark:/48223/pf0000381137>.

[361] Floridi, L., Cowls, J., Beltrametti, M., et al. (2018) *AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations*. *Minds & Machines* 28, 689–707.

between mandatory requirements and self-governance practices to safeguard people without hindering innovation.[362]

Although a "by design" approach is deemed crucial to ensure the effective safeguarding of human rights and ethical concerns in the digital realm,[363] it is increasingly clear that the design dimension is not limited to codes and digital architectures and that it should involve the social dimension of AI development, which includes governance and accountability mechanisms.[364]

In particular, the European Union has been active in developing a regulatory framework grounded in fundamental rights to position trustworthy and human-centric AI as the "distinctive trademark for Europe and its industry as a leader in cutting-edge AI"[365] and set the global standard for future AI. Since its 2019 Communication "Building Trust in Human-Centric Artificial Intelligence", the European Commission identified accountability as one of the key requirements that AI applications should respect to be considered trustworthy.

As stated:

"Mechanisms should be put in place to ensure responsibility and accountability for AI systems and their outcomes, both before and after their implementation. Auditability of AI systems is key in this regard, as the assessment of AI systems by internal and external auditors, and the availability of such evaluation reports, strongly contributes to the trustworthiness of the technology".[366] Currently, the European Union is discussing a proposal for a 'Regulation Laying Down Harmonised Rules On Artificial Intelligence', better known as

[362] Turner, J. (2019) *Robot Rules – Regulating Artificial Intelligence*. London: Palgrave. Brown, I., and Mardsen, C. (2013) *Regulating Code*. London: The MIT Press; Brownsword, R., and Yeung, K. (2008) *Regulating Technologies*, Oxford: Hart Publishing.

[363] Suzor, N., Dragiewicz, M., Harris, B., Gillett, R., Burgess, J., and Van Geelen, T. (2019). Human Rights by Design: The Responsibilities of Social Media Platforms to Address Gender-Based Violence Online. *Policy & Internet*, 11, 84–103. Cath, C., and Floridi, L. (2017) The Design of the Internet's Architecture by the Internet Engineering Task Force (IETF) and Human Rights. *Sci Eng Ethics*, 23, 449–468.

[364] Shilton, K. (2015). "That's Not an Architecture Problem!": Techniques and Challenges for Practicing Anticipatory Technology Ethics. In *iConference 2015 Proceedings 7*. iSchools.

[365] European Commission (2019) Communication 168 "Building Trust in Human-Centric Artificial Intelligence.", p.9.

[366] European Commission (2019) Communication 168 "Building Trust in Human-Centric Artificial Intelligence.", p.6.

'Artificial Intelligence Act' (AIA), establishing regulatory requirements for AI systems.

This paper aims to discuss accountability mechanisms in the AIA. After briefly introducing the concept of accountability in the artificial intelligence field, the next section will illustrate the institutional framework established by AIA to provide external accountability. Then, the identified external accountability mechanisms will be critically assessed.

While recognizing the relevance of the AIA as the first proposed regulation setting an accountability framework for the AI, the paper also warns about the risk that the intended purpose would be undermined by an institutional setting that it is not best placed to assess the social implications of technical specifications and solutions without exposing to special interests capture.

## 9.2 Accountability and Artificial Intelligence

Accountability can be understood as "a relationship in which a decision-maker is asked to report on their activities, and likely involving sanctions in the case of misconduct".[367]

Scholars usually distinguish between internal and external accountability. Internal accountability refers to a principal-agent relationship, in which an agent has been delegated to act on behalf of the principal and so it must report to the principal for his behavior, and it could be removed. This is typically the case of the board of directors against the shareholders of the company of society. External accountability requires agents to justify their behavior "to people or groups outside the acting entity who are nevertheless affected by it"[368] or in front of the broader general public.

As noted, within the realm of AI, accountability assumes a "networked configuration", in which "multiple actors have the obligation to explain and justify their use, design, and/or decisions of/concerning

[367] Palladino, N., & Santaniello, M. (2021). Legitimacy, power and inequalities in multistakeholder Internet governance. Cham: Palgrave McMillan, p.34.

[368] Risse, T. (2006). Transnational Governance and Legitimacy. In A. Benz, et al. (Eds.), Governance and Democracy Comparing National, European and International Experiences. New York: Routledge, p.185.

the system and the subsequent effects of that conduct,"[369] following the various stages of the system's lifecycle.

In this view, internal accountability in the artificial intelligence field is made problematic by the so-called many hands problem which refers to the fact that the development of AI systems involve different kind of actors at various stages.[370] So we can think about internal accountability like a chain of duty and responsibilities between different internal stakeholders. Instead, external accountability requires the evaluation of an external forum such as public opinion or authority such as a certification body. External accountability is particularly important in the context of artificial intelligence.

Since artificial intelligence system incorporate within their architecture social values, norms, assumptions about the nature of the world external accountability helps to draw out embodied values and requires decision makers to justify their choices and they algorithmic systems outputs in front of public reason, that means according to epistemic and normative standards which are acceptable to all reasonable people.[371]

Most of the international frameworks on trustworthy AI focus on a narrow set of requirement to achieve accountability. These requirements serve both internal and external accountability purposes. The two dimension indeed are strictly interrelated and mutual reinforcing. On the one side, internal accountability mechanisms provide the documentation necessary for a third party inspection and ensure the system's auditability by external parties. On the other side, external accountability duties compel all parties engaged in AI development, deployment, and management to meticulously document and justify their decisions while closely monitoring their outcomes.

More in detail, the aforementioned requirements consist of:

[369] Wieringa, M. (2020). What to account for when accounting for algorithms: A systematic literature review on algorithmic accountability. In Proceedings of the 2020 conference on fairness, accountability, and transparency, p.10.

[370] Schiff, D., Rakova, B., Ayesh, A., Fanti, A., & Lennon, M. (2020). *Principles to Practices for Responsible AI: Closing the Gap* (arXiv:2006.04707). arXiv. <http://arxiv.org/abs/2006.04707>.

[371] Binns, R. (2018). Algorithmic Accountability and Public Reason. *Philosophy & Technology*, *31*(4), 543–556. <https://doi.org/10.1007/s13347-017-0263-5>.

**(i)** Liability and Legal Responsibility: As exemplified in the Chinese AI Industry Code of Conduct, it is imperative to elucidate. “the rights and obligations of parties at each stage in research and development, design, manufacturing, operation, and service of AI, to be able to promptly determine the responsible parties when harm occurs.“[372]

**(ii)** Verification and Validation: AI systems’ providers must furnish proof that their application operates accurately in line with anticipated performance standards. Following the IEEE[373] “verification is a demonstration that a given application meets a narrowly defined requirement; validation is a demonstration that the application answers its real-world use case.”

**(iii)** Assessments: Before being placed on the market, AI systems must “be subjected to tests that do not put people’s lives in danger, harm their quality of life, or negatively impact their reputation or psychological integrity.”[374]

**(iv)** Auditability: AI systems must be designed to allow for third-party inspection. This means that “models, algorithms, data, and decisions should be recorded” to be inspected,[375] and access should be granted to competent supervisory authorities even through the support of proper interfaces.

**(v)** Appealability and Remediability: The determinations made by AI systems should be subject to dispute within relevant entities, and mechanisms for addressing negative consequences should be put in place.[376]

The Artificial Intelligence Act is the first framework attempting to institutionalize the aforementioned dimensions of accountability. Subsequent paragraphs will undertake a more comprehensive exploration of the mechanisms drafted in the proposal to ensure accountability.

[372] <https://www.secrss.com/articles/11099>.

[373] IEEE (2019). *Ethically Aligned Design* First Edition, <https://standards.ieee.org/industry-connections/ec/ead1e-infographic/>, p.269.

[374] <https://recherche.umontreal.ca/english/strategic-initiatives/montreal-declaration-for-a-responsible-ai/>.

[375] Association for Computing Machinery US Public Policy Council (USACM), (2017). *Statement on Algorithmic Transparency and Accountability*, <https://www.acm.org/binaries/content/assets/public-policy/2017_joint_statement_algorithms.pdf>, p.2.

[376] Amnesty International and Access Now (2018). *The Toronto Declaration: Protecting the right to equality and non-discrimination in machine learning systems*. <https://www.torontodeclaration.org/>.

## 9.3 Accountability mechanisms in the AIA

In the last few years, the European Union has turned its attention to AI regulation as a key policy issue (EU) to guarantee that AI systems are created and operate in accordance with EU values and principles promoting a "human-centric" approach to AI.

To this purpose, the Commission released the "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" in April 2021, best known as the Artificial Intelligence Act (AIA draft proposal). The proposal has been amended by the Council (December 2022) and the European Parliament (June 2023), and at the time of this writing it is undergoing the final phase of the legislative process, called "trilogue", in which the three institutions negotiate to produce an agreed version of the text.

This paragraph will first provide a brief overview of the approach and the main characteristics of the Artificial Intelligence Act draft proposal. Then, we will delve into the accountability mechanisms that have been foreseen in the Act.[377]

In its 2020 *White Paper on Artificial Intelligence,*[378] the European Commission outlined the need to develop an "ecosystem of trust" to foster the widespread adoption of AI, by addressing the potential risks associated with specific applications of this novel technology. Initially, the European Commission proposed a soft-law approach by releasing its non-binding 2019 Ethics Guidelines for Trustworthy AI. However, in 2021, a shift occurred toward a legislative approach[379] with the publication of the Communication on *Fostering a European Approach to Artificial Intelligence.*[380]

[377] In doing so, this paper will primarily refer to the Commission's version, while also considering changes proposed by the Council and the Parliament in the discussion section when relevant.

[378] European commission. (2020). *White Paper on Artificial Intelligence.* <https://commission.europa.eu/publications/white-paper-artificial-intelligence-european-approach-excellence-and-trust_en>.

[379] Tambiama M. (2022). Artificial Intelligence Act Briefing (European Parliamentary Research Service). European Parliament.

[380] European commission. (2020). *Fostering a European Approach to Artificial Intelligence.* <https://digital-strategy.ec.europa.eu/en/library/communication-fostering-european-approach-artificial-intelligence>.

Concerned that existing legislation, might not adequately address the risks posed by AI technologies in terms of safeguards of fundamental rights, safety and consumer protection, the Commission proposed the implementation of new rules governing the development, market placement, and utilization of AI systems. These latter would align with existing regulations on product safety and be introduced alongside a new Machinery Regulation aimed at adapting European safety standards to deal with emerging technological products.

The AIA draft proposal applies and establishes obligations for different categories of actors, involved in the development, deployment and management of AI systems, including AI providers, users, importers, distributors.[381] In this manner, the Act contributes to the realization of the 'networked configuration' of accountability throughout the entire AI lifecycle we mentioned earlier. However, considering the need for a separate discussion to analyze the different profiles of these subjects and their respective obligations, throughout the remaining text, I will use the label 'AI provider' as a general term referring to those individuals or entities that put AI applications to the market or make them available to the public.

The proposal adopts a risk-based approach distinguishing AI application posing: (i) unacceptable risk, (ii) high risk, (iii) limited risk or minimal risk, and tailoring differentiated regimes for the different risk categories.

AI applications posing unacceptable risks are those AI systems considered a clear threat to the safety, livelihoods and rights of people. This category includes social scoring by governments, applications deploying subliminal techniques or exploiting vulnerable groups; or real-time remote biometric identification systems in publicly accessible spaces. These AI applications are banned with few exceptions and must not be placed on the market or put into services or use in the EU.

The proposal considers high-risk applications those AI systems that are safety components of a product; or are themselves a product, or are required to undergo a third-party conformity assessment, subject to Union harmonization legislation listed in Annex II of the proposal.

[381] A more comprehensive description of these figures is available in Article 3 of the Act.

Moreover, the proposal identifies a series of high-risk applications in eight specific areas listed in Annex III, which could be updated as necessary by way of a delegated act (Article 7). By and large, all the applications that significantly harm the health, safety, and fundamental rights of persons could be considered high-risk applications and added to the list.

The proposal set out a range of requirements high-risk systems must comply with. According to the Act, they have to put in place a Risk Management System (Art.9), a Data Governance System (Art.10); a Record Keeping System (Art.12), a Quality Management System (Art.17), and a Post Market Monitoring System (Art.61). Moreover, high-risk AI systems shall be designed in such a way that they can be effectively overseen by natural persons during the period in which the AI system is in use (Human Oversight, Art.14), and they achieve appropriate levels of Accuracy, Robustness and Cybersecurity (Art.15). Furthermore, AI providers shall ensure proper Transparency and provision of information to users (Art.13) and they have to draw up the Technical Documentation (Art.11) concerning the measures undertaken to comply with the Act's requirements.

It is worth noting that the Act does not indicate specific implementable arrangements to comply with these requirements. Rather, it entrusts providers of AI systems with the task of identifying or developing solutions according to the most up-to-date and validated scientific knowledge and agreed-upon standards.

In December 2020, the European Commission published a standardization request addressed to the European Standardization Organizations (CEN, CELEC, ETSI) to develop a series of Harmonised Standards to comply with the Artificial Intelligence Act's requirements.

Harmonised Standards are standards specifically designed by a recognized European Standards Organisation to support EU legislation, following a request from the European Commission.

They are published in the Official Journal of the European Union (OJEU), and adhering to them carries a "presumption of conformity" with the essential requirements.

If the standard organizations decline the request or the harmonized standards are not delivered in time, the Commission will establish its own "common technical specifications" after consulting *ad hoc* expert groups.

AI applications not included in the previous categories are considered limited or minimal risk. They are subject to some mandatory requirements in the case they interact with humans (i.e., chatbots), emotion recognition systems, biometric categorisation systems, and AI systems that generate or manipulate image, audio or video content (i.e. deepfakes). In any case, the Commission encourages the drawing up of codes of conduct intended to foster the voluntary application of the requirements set out in the proposal.

The AIA proposal contains several provisions designed to ensure accountability.

With regard to the dimension of *Liability and Responsibility*, the Act mandates that high-risk AI systems must be registered in an EU-wide database managed by the Commission before they can be placed on the market or put into service (Art. 60). Additionally, AI providers are requested to have a legal representative on the territory of the EU. Furthermore, one of the provisions related to the Quality Management System is an accountability framework setting out the responsibilities of the management and other staff.

*Validation and verification* measures are requested to ensure a proper level of accuracy, robustness, and cybersecurity, as well as the quality of the dataset used to train and test the AI model.

Concerning the *Assessment* dimension, the Act, as said, foresees to put in place a Risk Management System, in order to identify potential risks and adopt proper measures to eliminate or mitigate them.

Several provisions in the draft proposal aim to ensure the *Auditability* of AI systems. For instance, the requested Technical Documentation should include a comprehensive description of the system architecture and its design process, encompassing hardware, software, and data components, along with their interactions and the human oversight mechanisms in place. Furthermore, the Record Keeping system should facilitate the automated recording of the states and operations of

AI applications to ensure a proper level of traceability in the AI system's functioning.

However, the proposal has faced criticism for lacking provisions related to *Appealability and Remediability*. Another piece of legislation currently under discussion will partially address this deficiency, the AI Liability Act, which aims to provide a simplified procedure for individuals harmed by AI applications to seek compensation. Nevertheless, the issue of how to address unfair or incorrect automated decisions still requires attention.

It is worth noting that most of these provisions converge into what could be defined as the primary accountability mechanism in the Act: the Conformity Assessment. This is the procedure through which high-risk AI application providers must demonstrate their compliance with the requirements set forth in the Act or existing product safety legislation. By doing so, they can obtain the CE mark, which allows for distribution throughout the EU.

According to AIA art.43, this process could follow a twofold path.

Where harmonized standards or common specification are available, AI application providers may opt for the conformity assessment procedure based on internal control, basically a self-certification in which the provider releases an 'EU declaration of conformity' stating that the high-risk AI system in question meets the requirements set out in the Act, inasmuch it has been developed in conformity with harmonized standards or community specifications build to this purpose.

In the case harmonized standards and common specifications do not exist yet, and in other specified and limited cases, the AI system shall go through a third-party audit, which is a conformity assessment procedure based on the assessment of the quality management system and the technical documentation, with the involvement of a so-called 'notified body.'

Notified bodies are conformity assessment bodies entitled by ad hoc 'notifying authority' designated or established by each member state.

In both cases, anyway, AI providers are subject to further external accountability mechanisms. First, we have the surveillance of designed market surveillance authorities (art.63), which are conferred with may powers, including the power to require relevant documents, technical specifications, data or information on compliance and technical aspects of the product, as well as the supply chain; the power to require corrective action in the case of non-compliance or emerging risks, the power to impose penalties or take other measures in the case AI provider fails to take appropriate corrective action or where the non-compliance or the risk persists, including the power to prohibit or restrict the making available of a product on the market or to order that the product is withdrawn or recalled.[382]

Moreover, under the AIA, AI applications are also subjected to the scrutiny of designated national public authorities or bodies supervising the respect of obligations under Union law protecting fundamental rights, which can access all the documentation produced under AIA provisions or request market surveillance authorities to organize specific tests (art.64).

Finally, the AIA foresees economic penalties for non-compliance with the prohibitions and requirements set in the regulation.

## 9.4 Discussion and Conclusion

The AI Act is emerging as a milestone in the history of AI governance. It will be the first comprehensive regulation on AI, establishing a framework for the development and use of AI within the European Union.

Furthermore the implementation of the Artificial Intelligence Act could set a precedent for other countries and regions to follow, in establishing their own regulations for the responsible development and use of AI, similar to what happened in the case of GDPR.

It would also be the first institutionalized accountability system for AI, outlining a framework specifying what requirements AI applications are expected to satisfy; what are the responsibilities of providers, distributors, importers and users of AI applications; who are the authorities enforcing the rules and what are the sanctions in case of misconduct.

[382] These are the powers assigned to market surveillance authorities by Regulation 2019/1020.

However, the AIA also raises some concerns, questioning the effectiveness of this instrument.

Setting a series of technical and organizational requirements for the development, deployment, and management of AI systems is undoubtedly a crucial step in the achievement of a human-centric and trustworthy AI.

Nevertheless, in an effort to not pose excessive burdens on the shoulders of the nascent European AI industry and, in so doing, hinder the EU's geopolitical ambition to be a global player in the field, AIA requirements apply only to high-risk applications. Moreover, as seen, when harmonized standards or common specifications are available, compliance with these requirements will be ascertained mostly through a self-assessment procedure.

This means that most AI applications that will be released in the European market in the coming years will not undergo any prior evaluation by third-party bodies of the presence and adequacy of measures to ensure their safety and respect for fundamental rights, including the vast majority of applications classified as high-risk under the same regulation.

Along with the limitedness of the application and preventive check of the requirements established in the act, another, and probably most relevant issue consists of the capability of EU institutions to evaluate the adequateness of the technical means put in place to comply with the established requirements in an autonomous manner and avoiding special and vested interests capture.

This relates to at least two different points.

**a.** First, we have the evaluation of the harmonized standard. The standard-setting bodies entrusted to develop the harmonized standard are non-profit private organizations in which companies could exert a notable influence[383] and attempt to soften the burdens on businesses.

[383] Palladino, N., & Santaniello, M. (2021). *Legitimacy, power and inequalities in multistakeholder Internet governance*. Cham: Palgrave McMillan, Van Klyton, A., Arrieta-Paredes, M.-P., Palladino, N., & Soomaree, A. (2023). Hegemonic practices in multistakeholder Internet governance: Participatory evangelism, quiet politics, and glorification of status quo at ICANN meetings. *The Information Society*, *39*(3), 141–157. <https://doi.org/10.1080/01972243.2023.2194295>.

EU regulation 1025/2012 entrusts the Commission with the responsibility to decide about the adequateness of the proposed standard, with the assistance of an *ad hoc* committee or other group of experts. Besides this is a well-established procedure with a proven track record of successful cases, there are reasons that suggest exercising caution.

Other complex technologies, such as nuclear power or biotech, can have relevant and tangible impacts on people's safety and other remarkable social implications, but typically, they are limited in scope, affecting specific sectors or areas of human experience.

Instead, AI is an extremely pervasive technology that can affect many different aspects of our lives, being employed in every sector of social life, from health to economy, leisure, communication, and so on.

Moreover, the impact of AI can be more deep-seated and subtle. Artificial intelligence are systems that make decisions with a certain degree of autonomy learning by the interaction with their environment. In so doing, they influence the way in which decisions are taken in our society, which also means that they change the way in which our society, our states, and companies are organized, they influence the way in which we interact and behave, and even our identity. AI applications are involved in decision-making processes that can determine if we will be hired, or if we can have a loan or access to a university. Social media platforms' algorithms can influence our opinions, as well as our mood or self-esteem.

In other words, AI architectures can become governance architecture, giving rise to a digital infrastructural governance layer capable of disciplining human behavior.

For these reasons, the evaluation of AI standards cannot rely merely on technical considerations about efficiency but should be grounded on a deep understanding of the social implications of technical specifications and the capability to translate political aims into socio-technical architecture.

In this case, on the one hand, the Commission lacks the internal competencies and expertise required to fully comprehend the implications of the specific organizational and technical arrangements and tools proposed in the harmonized standards. On the other hand, within the field of AI, experts possessing this level of understanding are often associated with the business sector or have significant ties with private companies, much like many academics. This dynamic could potentially compromise the impartial consideration of public interests during the assessment of harmonized standards, potentially leading to the dominance of specific interests.

**b.** Similar considerations could be advanced in relation to market surveillance authorities, which have a pivotal role in the oversight of the compliance with the regulation once AI systems have been released on the market. Even in this case, these institutions are expected to possess the necessary technical expertise to carry out the designated responsibilities, as stipulated by Regulation 2019/1020 concerning market surveillance.

However, market surveillance authorities are typically integrated within ministries and authorities, which are unlikely to already have the required personnel and structures in place.

It may be that specific recruitment or *ad hoc* structures will be established to align with the intended objectives. Nevertheless, considering the extensive proliferation of AI technologies and the resulting vast number of AI applications across diverse sectors and contexts, each with its distinct characteristics, concerns arise regarding the capacity of smaller entities, such as ministerial or already existing authority departments to manage these responsibilities effectively.

To address these challenges, a way forward should involve enhancing expertise; ensuring impartiality, and establishing robust supervisory authority.

With regard to the first point, the European Commission and other relevant bodies should invest in building internal expertise and understanding the societal implications of AI standards. This includes recruiting experts with a deep understanding of both the technical and social aspects of AI.

Coming to impartiality, mechanisms should be put in place to prevent conflicts of interest among experts involved in evaluating AI standards.

Finally, EU decision-makers should consider creating larger supervisory authorities specialized in AI governance, similar to data protection supervisory bodies. These authorities could pool resources and expertise to effectively oversee AI applications across various sectors.

However, we should recognize that the challenges in regulating AI are multifaceted and may require innovative solutions. Public authorities should continuously adapt their structures and processes to address the challenge to aligning organizational and technical solutions to comply with the requisites of human-centric and trustworthy AI within the unique parameters of various sectors and contexts.

# 10 Promoting the Transparency of AI-Generated Inferences

**Attamongkol ("Atta") Tantratian, Doctor of Juridical Science candidate, Indiana University Maurer School of Law, USA**

## Abstract

Many businesses today use artificial intelligence to generate inferences from the personal data they collect. While this practice enables companies to better understand individual consumer preferences, it often occurs in secret, without consumers' awareness. As a result, consumers are typically unaware of the inferences generated about themselves, compromising their privacy and autonomy as they lack control over the flow and use of such data. Despite the existence of data subject rights within current data protection regulations, such as the General Data Protection Regulation, access to inferences about themselves is not guaranteed. Businesses can deny inferences access requests by exploiting the broad scope of the law of trade secrecy: to cite their interest in protecting such data as trade secrets. Therefore, this Article deems it essential to reexamine the scope and application of trade secrets law in this context. It offers a descriptive analysis of the underlying legal frameworks in the U.S. and EU that empower businesses to classify consumer inferences as trade secrets. In response, the Article proposes that data protection or consumer protection authorities should carefully examine the scope of trade secrets law in their respective jurisdictions and issue guidelines to limit potential abuse of the law.

## 10.1 Introduction

> "You know BASH has over 40 million data points on you on every decision you have made since 1995. My algorithms have determined 8 fundamental consumer profile types. You are a Lifestyle Idealist. To 96.5% accuracy, your death was so unremarkable and boring. You're gonna die alone"[384]

[384] McKay, A. (2022). Don't Look Up [Motion Picture].

Using advanced algorithms to generate intimate inferences is no longer fictional. In today's personal-data-driven economy, either described as *Surveillance Capitalism*,[385] *Informational Capitalism*,[386] or *Inference Economy,*[387] consumer data is not merely collected and used as is. Instead, consumer data is processed using advanced algorithms to draw inferences that can help businesses generate more profits and make decisions on the data subjects.[388]

To conduct personalization for behavioral advertisements, for instance, businesses utilize artificial intelligence (AI) tools to analyze data collected from consumers and generate a wide range of inferences about them, from simple attributes to sensitive information such physical and mental health, and religious and political beliefs.[389] These inferences are then assigned to individuals in various forms, such as scores, tags, or categories.[390]

One grocery store chain in the U.S., Target, for instance, was revealed in 2012 to have secretly generated inferences about customers in a form of a pregnancy score based on product purchase history and other collected personal data. These scores were accurate enough that Target was able to identify a teenager's pregnancy even before her father did.[391] Another notable instance is the infamous Facebook Cambridge Analytica scandal, in which Facebook users categorized into different political profiles based on their quiz answers and

[385] Surveillance capitalism refers to "a new economy order that claims human experience as free raw material for hidden commercial practices of extraction, prediction, and sales. See Zuboff, S. (2019). *The age of surveillance capitalism: The fight for a human future at the new frontier of power* (chapter Definition).

[386] Informational capitalism refers to a political and economic model that focuses on extracting value from data. Cohen, J. E. (2019). *Between truth and power* (pp. 5-6). Oxford University Press.

[387] In the inference economy, "organizations use available data collected from individuals to generate further information about both those individuals and about other people." Solow-Niederman, A. (2022). Information privacy and the inference economy. *Northwestern University Law Review, 117*, 361.

[388] Pasquale, F. (2015). *The black box society* (pp. 25-32). Harvard University Press (discussing the scored society in which secret algorithms make inferences about people to profile, sort, and penalize them).

[389] Wachter, S. (2020). Affinity profiling and discrimination by association in online behavioral advertising. *Berkeley Tech Law Journal, 35*, 376-377.

[390] Solove, D. J. (2004). *The digital person: Technology and privacy in the information age* (p. 46). New York University Press ("... database marketers frequently classify consumers into certain categories based on stereotypes about their values, lifestyle, and purchase habits.").

[391] Duhigg, C. (2012, February 16). How companies learn your secrets. *The New York Times*. Retrieved from <https://www.nytimes.com/2012/02/19/magazine/shopping-habits.html?searchResultPosition=1>.

platform-monitored data, including "Like" clicks, and time spent on each post.[392]

For consumers, on the other hand, they often lack awareness about the inferences generated about them and, as a result, have no control in the process.[393] In 2017, a French user of the popular dating platform Tinder requested a copy of her personal data held by the company. While Tinder provided her with 800 pages of the raw data it had *collected* from her, it refused to disclose the inferences it *generated* (referred to by the user as "dark secrets") about her and how they were used in potential matchmaking processes. Tinder justified its decision by stating that such data and its matchmaking tools are in the "core part of our technology and intellectual property."[394] In 2018, Tinder maintained the same stance in response to a similar request from another user.[395] Similarly, back in 2011, Facebook rejected an access request made by privacy activist Max Schrems. Facebook's letter stated: "Section 4(12) of the [Irish Data Protection] Act carves out an exception to subject access requests where the disclosures in response would adversely affect trade secrets or intellectual property."[396]

This Article argues that it is essential to reexamine the scope of trade secrets in relation to data subjects' access requests (DSARs). While existing literature often discusses the application of trade secrecy to AI algorithms,[397] this Article focuses on the transparency of AI

[392] Weiss, B. (2018, March 18). Trump-linked firm cambridge analytica collected personal information from 50 million Facebook users without permission. *Business Insider*. Retrieved from <https://www.businessinsider.com/cambridge-analytica-trump-firm-facebook-data-50-million-users-2018-3>; Ma, A., & Gilbert, B. (2019, August 23). Facebook understood how dangerous the Trump-linked data firm Cambridge Analytica could be much earlier than it previously said. Here's everything that's happened up until now. *Business Insider*. Retrieved from <https://www.businessinsider.com/cambridge-analytica-a-guide-to-the-trump-linked-data-firm-that-harvested-50-million-facebook-profiles-2018-3#was-it-legal-4>.

[393] Marks, M. (2021). Emergent medical data: Health information inferred by artificial intelligence. *U.C. Irvine Law Review, 11*, 1000 (highlighting that inferences drawn from the collected personal data are often without consumers' knowledge or consent).

[394] Duportail, J. (2017, September 26). I asked Tinder for my data. It sent me 800 pages of my deepest, darkest secrets. *The Guardian*. Retrieved from <https://www.theguardian.com/technology/2017/sep/26/tinder-personal-data-dating-app-messages-hacked-sold>.

[395] Schmid, J. (2019, August 13). My GDPR complaint against tinder [web blog comment]. Retrieved from <https://medium.com/personaldata-io/my-gdpr-complaint-against-mtch-technology-services-139087d3de8a>.

[396] For a copy of the letter Facebook sent to Schrems, visit <http://www.europe-v-facebook.org/FB_E-Mails_28_9_1.pdf> (accessed Jul. 6, 2023).

[397] See for example: Lu, S. (2021). Algorithmic opacity, private accountability, and corporate social disclosure in the age of artificial intelligence, *Vanderbilt Law Review, 23*, 99 (suggesting algorithmic disclosure under securities law to promote transparency in Big Data.); Eaglin, J. M. (2017).

products (i.e., inferences), a topic that receives less attention. It assumes that inference transparency could equip ordinary consumers with the knowledge needed to make informed decisions about exercising their rights (e.g., to edit and delete). On the other hand, algorithmic transparency may not fulfill that purpose due to the inherent complexity of algorithmic models generally.[398]

Part II provides an overview of the underlying legal frameworks in the EU and the U.S. that enable businesses to withhold disclosure of inferences from consumers. Part III explores the potential harms inferences and the importance of advocating for data subjects' ability to access inferences about themselves. Part IV presents a framework for challenging data controllers' trade secret claims, based on the literature of U.S. trade secrets law. Data protection authorities across jurisdictions may adapt and further explore for implementation.

## 10.2 Cause of Inference Secrecy: The Underlying Legal Framework

The problem of inference opacity underscores a larger – and a global – concern: the asymmetry of power and knowledge between consumers and businesses. As noted by Amy Kapczynski, its root cause is not a lack of market regulations, but rather the impact of existing regulations, including trade secrets law.[399] This Article further establishes that alongside trade secrecy law, data privacy regulations also contribute to the issue, as the latter often prioritizes businesses' interests in protecting trade secrets over data subjects' rights.

## 10.3 Trade Secrecy Law

In defining trade secrets, the World Trade Organization Agreement on Trade-Related Aspects of Intellectual Property ("TRIPS Agreement") recognizes a trade secret as any information that (i) is secret or

Constructing recidivism risk. *Emory Law Journal, 67*, 59 (arguing the use of AI in criminal justice should be accountable and transparent); Wachter, S., Mittelstadt, B., & Floridi, L. (2017). Why a right to explanation of automated decision-making does not exist in the general data protection regulation. *International Data Privacy Law, 7(2)*, 85-89. <https://doi.org/10.1093/idpl/ipx005 (finding trade secrets have formed a barrier to algorithmic transparency, accountability, and fairness).

[398] Perl M. & Elkin-Koren N. (2017). Black box tinkering: Beyond disclosure in algorithmic enforcement, *Florida Law Review, 69*, 186-90 (arguing algorithmic disclosure would hardly help users understand the logics of automated decision making).

[399] Kapczynski, A. (2020). The law of informational capitalism. *Yale Law Journal, 129*, 1501.

not known to the public, (ii) has commercial values thanks to its secrecy, and (iii) has been reasonably kept secret.[400] In the U.S., trade secrecy has long been recognized,[401] covering "all forms and types of financial, business, scientific, technical, economic, or engineering information"[402] as well as "a formula, pattern, compilation, program, device, method, technique, or process."[403]

As a result of the broad definition and protection of trade secrecy law, data-driven companies, such as Myriad Genetics and Google,[404] can protect not only their data analytics tools but also the data generated from those tools as trade secrets.[405] Provided that the three factors (i.e., not publicly known, valuable thanks to secrecy, and kept confidential) are satisfied, personal data and inferences drawn about individuals—including shopping habits, profiles, creditworthiness, lifestyle, reliability, estimated life span, and work advancement—may as well fall under the expansive scope of protection of trade secrecy.[406]

## 10.4 Data Privacy Law

Furthermore, the regulations currently in effect in both the EU and U.S. tend to prioritize safeguarding businesses' trade secrets and the intellectual property (IP) over rights of data subjects in conflict.

The EU General Data Protection Regulation (GDPR) acknowledges the right of access to personal data and relevant information.[407] While not explicitly stated, there is less debate now about inferences related to an individual being considered personal data under the GDPR. In 2022, the European Data Protection Board (EDPB) released a set of

[400] Agreement on Trade-Related Aspects of Intellectual Property (TRIPS) art. 39.

[401] See for example: Vickery v. Welch, 36 Mass. (19 Pick.) 523, 525 (1837).

[402] Defend Trade Secrets Act of 2016, 18 U.S.C. § 1839(3) (2018).

[403] Uniform Trade Secrets Act § 1(4).

[404] See generally Simon, B. M., & Sichelman T. (2017). Data-generating patents. *Northwestern University Law Review, 111*, 377.

[405] Mattioli, M. (2014). Disclosing big data. *Minnesota Law Review, 99*, p. 556 (observing that business practices in Big Data are capable of being protected as trade secrets).

[406] Wachter, S. & Mittelstadt, B. (2019). A right to reasonable inferences: Re-thinking data protection law in the age of big data and AI. *Columbia Business Law Review, 2019*, 607 (noting these sample subject matters in the context of the application of EU Trade Secrets Directive). These examples, although are given under the context of EU law, could also be valid in the United States as both the United States and the European Union are members of the TRIPS Agreement.

[407] Directive 95/46/EC (General Data Protection Regulation) article 15.

guidelines confirming that *observed*, *derived*, and *inferred* data are personal data that must be disclosed to a data subject upon request.[408]

Furthermore, in a case from that year, the European Court of Justice (CJEU) ruled that inferences suggesting sensitive attributes (e.g., political opinions, trade union membership, and sexual orientation) are subject to the same rules as special categories of data under the GDPR.[409] As a result, Michael Veale and other experts predicted that in most cases, generating sensitive inferences would trigger the requirement for obtaining explicit consent.[410] Ideally, the requirement would help keep the data subjects informed and retain their control.

However, because the notice and consent process typically happens before data processing, data subjects might not be informed about subsequently generated sensitive inferences. In such cases, the right of access would serve as the remedy. Nonetheless, since the GDPR limits the right of access by stating that the right "shall not adversely affect the rights and freedoms of others,"[411] including "trade secrets or intellectual property,"[412] the 2022 EDPB guidelines affirm that data controllers may reject a DSAR if granting it would compromise their trade secrets rights.[413]

In the U.S., notable examples of state privacy laws include the California Consumer Privacy Act and Colorado Privacy Act Similar to the GDPR, the California law grants the right of access to personal data.[414] What sets it apart is its explicit inclusion of inferences as one category of personal data,[415] making it the first U.S. privacy law to do so.[416] Likewise, the Colorado law grants the right of access,

[408] The European Data Protection Board. (Feb. 1, 2022). Guidelines 01/2022 on data subject rights – right of access. 31. Retrieved from <https://edpb.europa.eu/our-work-tools/documents/public-consultations/2022/guidelines-012022-data-subject-rights-right_en>.

[409] OT v. Vyriausioji tarnybinės etikos komisija, Case 184/20 (CJEU Aug. 1, 2022).

[410] Lomas, N. (2022, August 2). Sensitive data ruling by Europe's top court could force broad privacy reboot. Retrieved from <https://techcrunch.com/2022/08/02/cjeu-sensitive-data-case/>.

[411] Directive 95/46/EC (General Data Protection Regulation) article 15(4).

[412] Directive 95/46/EC (General Data Protection Regulation) recital 63.

[413] The European Data Protection Board. (n.25) 49-50.

[414] California Civil Code §1798.110(a)(1).

[415] California Civil Code §1798.140(o)(1)(K).

[416] Blanke, J. (2020). Protection for 'inferences drawn:' A comparison between the General Data Protection Rule and the California Consumer Privacy Act. *Global Privacy Law Review, 2*, 92.

encompassing inferences ,[417] and go even further than the California law by providing a definition of "sensitive data inferences" with higher protection.[418]

Nevertheless, both laws contain exceptions for trade secrets rights. While the Colorado law states the exceptions explicitly,[419] the California law incorporates provisions for potential future exceptions, particularly concerning trade secrets and intellectual property rights.[420] Accordingly, the California State Attorney General noted in 2022 that while inferences are subject to disclosure under the consumers' right of access, businesses are not obligated to disclose inferences that qualify as trade secrets. [421] The prioritization of trade secrets of businesses over consumers' personal data rights is also evident in other recent state privacy legislations, including those of Tennessee, Iowa, Connecticut, Indiana, Montana, Oregon, Texas, and Utah.[422]

## 10.5 Data Subjects' Interest to Access Inferences about Themselves

### 10.5.1 Sophisticated Inference Generation

As the use of machine learning to process personal data becomes more common due to lower costs, businesses can now generate more inferences than before.[423] Further, with powerful machine learning algorithms today, Daniel Solove argues that sensitive inferences can be drawn from nearly all personal data.[424] As a result, consumers might be more aware of the data they disclose to a business (such as photos and date of birth) than the inferences generated by the business (such

[417] Colorado Privacy Act Rules 4 CCR-904-3, Rule 4.04 (Sept. 29, 2022).

[418] Colorado Privacy Act Rules 4 CCR-904-3, Rule 2.02 (Sept. 29, 2022).

[419] Colorado Privacy Act Rules 4 CCR-904-3, Rule 4.07(B) ("[A] Controller is not required to provide Personal Data to a Consumer in a manner that would disclose the Controller's trade secrets").

[420] California Civil Code §1798.185.

[421] Office of the Attorney General State of California. (Mar. 10, 2022). Opinion No. 20-303, 14-15.

[422] For each state bill, see IAPP U.S. State Privacy Legislation Tracker 2023, <https://iapp.org/media/pdf/resource_center/State_Comp_Privacy_Law_Chart.pdf>.

[423] Uberti, D. (2022, January 4). Come the metaverse, can privacy exist? *Wall Street Journal*. Retrieved from <https://www-wsj-com.cdn.ampproject.org/c/s/www.wsj.com/amp/articles/come-the-metaverse-can-privacy-exist-11641292206> (suggesting with the advancement of technologies, "now the data is in inferences").

[424] Solove, D. (2023). Data is what data does: Regulating use, harm, and risk instead of sensitive data. *Northwestern University Law Review, 118 (Forthcoming 2024)*, 5. Retrieved from <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4322198>.

as sexual orientation and political beliefs).[425] Consider data collection in online communications as an example. The table below illustrates possible inferences that could be drawn from seemingly unrelated sets of given data.

| Input data (collected personal data) | Output data (i.e., inferences) |
|---|---|
| Smartphone usage:<br>Calls, texts, and apps usage | Big-Five personality traits: extraversion, agreeableness, conscientiousness, neuroticism, and openness to experience.[426] |
| Facebook likes | Ethnic affinity.[427] |
| | Sexual orientation, ethnicity, religious and political views, personality traits, intelligence, happiness, use of addictive substances, parental separation, age, and gender.[428] |
| Language on social media | Big-Five personality traits.[429] |
| Search query histories. | Age, gender, political and religious views.[430] |
| Group photos | The importance of each person.[431] |
| Rhythm of their typing patterns on a standard keyboard | Emotional state[432] |
| Location tweets | Neighborhoods of users, which then reveal average income, average housing cost, debt, and other demographic information, such as political views.[433] |

[425] Wachter, S. (n.6). 376-377 (finding that data processors can generate inferences and predictions about "religious or political beliefs, sexual orientation, race or ethnicity, physical or mental health status, or sex or gender identity from online behavior without users ever being aware.").

[426] Chittaranjan G., Blom J., & Gatica-Perez, D. (2011). Who's who with big-five: Analyzing and classifying personality traits with smartphones. In Wearable Computers (ISWC), *2011 15th Annual International Symposium*, 29-36. <https://doi.org/10.1109/ISWC.2011.29>.

[427] Angwin, J., & Parris, T. (2016). Facebook lets advertisers exclude users by race. Retrieved from <https://www.propublica.org/article/facebook-lets-advertisers-exclude-users-by-race>.

[428] Kosinski, M., D. Stillwell, & T. Graepel. (2013). Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences, 110(15)*, 5802-5805. <https://doi.org/10.1073/pnas.1218772110>.

[429] Park, G., Schwartz, H.A., Eichstaedt, J.C., Kern, M.L., Kosinski, M., Stillwell, D.J., Ungar, L.H., & Seligman, M.E. (2015). Automatic personality assessment through social media language*. Journal of Personality and Social Psychology, 108(6)*, 934. <https://doi.org/10.1037/pspp0000020>.

[430] Bi, B., Shokouhi, M., Kosinski, M. & Graepel, T. (2013). Inferring the demographics of search users: Social data meets search queries. In Proceedings of the 22nd international conference on World Wide Web, 131-140. <https://doi.org/10.1145/2488388.2488401>.

[431] Mathialagan, C., Gallagher, A.C. & Batra, D. (2015). Vip: Finding important people in images. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 4858-4866. Retrieved from <https://research.google/pubs/pub43844/>.

[432] Epp, C., Lippold, M., & Mandryk, R.L. (2011) Identifying emotional states using keystroke dynamics. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, 715-724. <https://doi.org/10.1145/1978942.1979046>.

[433] Liccardi, I., Abdul-Rahman, A., & Chen, M. (2016). I know where you live: Inferring details of people's lives by visualizing publicly shared location data. In Proceedings of the 2016 Conference on Human Factors in Computing Systems, 1-12, <https://doi.org/10.1145/2858036.2858272>.

As social media platforms are increasingly providing more immersive experiences to users, the popularity of extended reality (XR) devices is on the rise. These devices come equipped with sensors that enable platform companies to collect a greater amount of user biometric data than previously. As a result, the potential for intricate and sensitive inference generation has also increased.[434] Below are examples that illustrate how sensitive inferences can be generated from biometrics data captured through XR devices.

| **Input data (collected by XR sensors)** | **Output data (i.e., inferences)** |
|---|---|
| Eye movements | Health status (such as autism, schizophrenia, Parkinson's, ADHD, and concussions), emotions, sexual interest, and inner thoughts. [435] |
| Gaze patterns | Biometric identity, gender, age, ethnicity, body weight, personality traits, drug consumption habits, emotional state, skills and abilities, fears, interests, sexual preferences, physical and mental health conditions. [436] |
| Behavioral data (e.g., reaction time, voice, vision, fitness) | Age and disabilities.[437] |
| Device data (e.g., CPU power, resolution, tracking and refresh rates) | Wealth.[438] |
| Geospatial data (e.g., height, left & right arms, room) | Wingspan, gender, room area.[439] |

Furthermore, in the *Inference Economy*, Alicia Solow-Niederman suggests that data from one individual may also be used to make predictions about other individuals.[440] This implies that data about

[434] Berrick D., & Spivack J. (Nov.17, 2022). Understanding extended reality technology & data flows: Privacy and data protection risks and mitigation strategies. Retrieved from <https://fpf.org/blog/understanding-extended-reality-technology-data-flows-privacy-and-data-protection-risks-and-mitigation-strategies/> ("XR devices collect, process, and share large quantities of data about users' bodies and environments. This data could be used to make inferences—whether accurate or not—about sensitive aspects of peoples' lives, such as their sexual orientation or health conditions.").

[435] Bar-Zeev, A. (May 28, 2019). The eyes are the prize: Eye-tracking technology is advertising's holy grail. Retrieved from <https://www.vice.com/en/article/bj9ygv/the-eyes-are-the-prize-eye-tracking-technology-is-advertisings-holy-grail>.

[436] Kröger, J. L., Lutz, O. H.-M., & Müller, F. (2020). What does your gaze reveal about you? On the Privacy Implications of Eye Tracking. Privacy and Identity Management. Data for Better Living: AI and Privacy, 226-241. <https://doi.org/10.1007/978-3-030-42504-3_15>.

[437] Nair, V., Gonzalo Munilla Garrido, & Song, D. (2022). Exploring the unprecedented privacy risks of the Metaverse. 23rd Privacy Enhancing Technologies Symposium, 238-256. <https://doi.org/10.48550/arxiv.2207.13176>.

[438] ibid.

[439] ibid.

[440] Solow-Niederman, A. (n.4).

one person can be used to make informed guesses about others, potentially revealing sensitive information they have not shared and may not wish to disclose.[441] As a result, all the potential inferences discussed in the tables above could also apply to and be associated with several other individuals besides the data subjects themselves.

## 10.6 The Case for Accessing Inferences

Despite the potential benefits that inferences can offer (e.g., positive nudges,[442] enhanced service experiences),[443] inference generation often raises concerns about privacy and autonomy. For some individuals, merely making inferences about someone alone can impact their ability and entitlement to self-identity. Commentators argue that privacy grants individuals with the right to determine their own identities and beliefs.[444] Inferences and predictions generated by businesses thus take away some of the consumers' ability to privately identify themselves. Additionally, certain inferential analyses blur the boundary between what individuals choose to reveal and what they choose to keep inside their minds. Inference generation thus implicates "the right to keep our thoughts and opinions private". According to Susie Alegre, such right is a key element of the internationally protected right to freedom of thought and is crucial in the digital age.[445]

From a more materialized perspective, uses of inferences could easily lead to the risk of discrimination and harm. Notably, many online publishers are known for allowing and supporting advertisers to display advertisements based on users' protected profiles. In the U.S., this

[441] ibis., 385.

[442] Möhlmann, M. (Apr. 22, 2021). Algorithmic nudges don't have to be unethical. *Harvard Business Review*. Retrieved from <https://hbr.org/2021/04/algorithmic-nudges-dont-have-to-be-unethical> (suggesting inferences can empower positive nudges).

[443] Personalization-privacy paradox: Why solving it matters now. (2022, July 4). Retrieved from <https://www.cdotrends.com/story/16573/personalization-privacy-paradox-why-solving-it-matters-now?refresh=auto> (finding personalized experience is what many consumers enjoy and expect from businesses, despite increasing privacy concerns).

[444] See for example: Richards, N. (2011). *Why privacy matters* (p. 113). Oxford University Press. ("Being able to say who I am—what my name is—is a basic prerequisite for being human, one that has long been appreciated in Western culture."); Allen, A. L. (1999). Coercing privacy. *William & Mary Law Review, 40*, 738 ("Privacy has value relative to normative conceptions of spiritual personality, political freedom, health and welfare, human dignity, and autonomy.").

[445] Alegre S. (2021). Protecting freedom of thought in the digital age. *Center of International Governance Innovation, 165*, 3-4. Retrieved from <https://www.cigionline.org/static/documents/PB_no.165.pdf.>

practice has resulted in discriminatory online behavioral advertising instances, such as when black people's names searches show more advertisements about arrest search websites, or job search engines suggest more STEM positions to male users compared to female users.[446] Also, large-scale inference generation often leads to inaccuracies,[447] with corrections being rarely possible.[448] These erroneous inferences could easily become entangled with all the other data about the individual, making it challenging to identify their source, accuracy, and impact.[449] As a consequence, these false inferences could potentially lead to a future unforeseeable harm for the individual.[450]

As data subjects struggle to predict the extent of information held by a data controller and its potential implications, they lack the information needed to effectively use other data protection remedies—such as correcting or deleting their personal data. Thus, commentators contend that the right of access, as a cornerstone for other rights,[451] should apply to inferences which the data subjects are typically less aware of.[452] In alignment with this view, the California State Attorney General, for instance, stated in 2022 that "inferences appear to be at the heart of the problem that the CCPA seeks to address."[453]

[446] Wachter, S. (n.6) 377-378.

[447] Cyphers, B., & Gebart, G. (2019). Behind the one-way mirror: A deep dive into the technology of corporate surveillance, 5-6. Retrieved from <https://www.eff.org/files/2019/12/11/behind_the_one-way_mirror-a_deep_dive_into_the_technology_of_corporate_surveillance.pdf> (suggesting that companies frequently use this data to derive conclusions that are inaccurate based on "better than nothing" statistical guesses.); Tene, O., & Polonetsky, J. (2013). Big data for all: Privacy and user control in the age of analytics. *Northwestern Journal of Technology & Intellectual Property, 11(5)*, 270; Solove, D. J. (n.7) p. 46 ("Not only are our digital biographies reductive, but they are often inaccurate.").

[448] Epp, C., Lippold, M., & Mandryk, R.L. (2011) Identifying emotional states using keystroke dynamics. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, 715-724. <https://doi.org/10.1145/1978942.1979046>.

[449] Blanke, J. (n.33) pp. 84-85.

[450] Cofone, I. (2022). Privacy standing. *Illinois Law Review, 2022*, p. 1384 (highlighting that "harmful information is rarely collected information and is frequently inferred information").

[451] Ausloos, J., Veale, M., & Mahieu R. (2020) Getting data subject rights right. *Journal of Intellectual Property, Information Technology, and Electronic Commerce Law*, para. 41. <https://doi.org/10.31228/osf.io/e2thg> (suggesting that "[a]ccess rights are the pre-requisites to so many other potential applicable rights…").

[452] Shah S. (Jan. 30, 2019). This lawyer believes GDPR is failing to protect you: Here's what we should change. Retrieved from <https://www.forbes.com/sites/soorajshah/2019/01/ 30/this-lawyer-believes-gdpr-is-failing-to-protect-you-heres-what-she-would-change/?sh=67141e596fc4> (suggesting that "the GDPR focuses too much on the input stage, meaning when data is collected, but not enough on how it is assessed. Once the data is lawfully obtained we have very little control or understanding of what inferences can be drawn").

[453] Office of the Attorney General State of California. (Mar. 10, 2022). Opinion No. 20-303, p. 13. Retrieved from <https://oag.ca.gov/system/files/opinions/pdfs/20-303.pdf>.

To safeguard the data subjects against potential harms from uses of inferences, legal scholars have argued for the *Right to Reasonable Inferences* to regulate high-risk inferences (i.e., causing potential reputation damage) that are not verifiable in nature.[454] This proposed right would impose obligations on businesses, before processing collected data, to demonstrate that a) it is normatively acceptable to draw inferences from the given data, b) the purposes are appropriate, and c) the processing methods and models are reliable. Also, the data subjects would have the ability to challenge the generated inferences that are "inaccurate or unreasonable." Nevertheless, the scholars noted that trade secrets law is a barrier to their proposed right and did not propose a solution to the problem.[455] Thus, without challenging the application of trade secrets law, the inference opacity problem would remain.

## 10.7 Challenging Trade Secrecy for Inference Transparency

Therefore, this Article takes the approach of questioning the law of trade secrecy itself, which the Article believes is the root cause of the problem. It examines the literature of trade secrets law in the U.S. and argues that some of the forgotten definitions, if brought back, could help break the rigid trade secrets barrier restricting inference access.

The Article thus urges data protection authorities (or consumer protection agencies, depending on jurisdictions) to critically examine their respective trade secrets law. This evaluation could lead to the issuance of comments or guidelines, facilitating data subjects' access to their inferences.[456] Specifically, the authorities should challenge the inherent nature and scope of trade secrets law in their jurisdiction, as well as impose documentation requirements to data controllers.

[454] Wachter, S. & Mittelstadt, B. (n. 23) 494-591.

[455] ibid., 606-610.

[456] This Article is of the opinion that they are entitled to comment on the scope of trade secrets protection when in conflict with data subjects' rights. In the U.S. for instance, no authority is primarily overseeing trade secrets, not even the United States Patent and Trademark Office (PTO). See for example: Goldman, E. (2016). The Defend Trade Secrets Act isn't an "intellectual property" law, *Santa Clara High Tech, 33*, p. 548 ("So where Congress authorizes the PTO to comment or advice on intellectual property, in theory the [Defend Trade Secrets Act] is not included"). The discussion of trade secret law also receives scant attention in the literature.

### 10.7.1 Examine the nature and scope of trade secrets law in the jurisdiction

A data protection authority may begin by examining the nature of trade secrets law in their jurisdiction to determine how trade secrets rights ought to be exercised. While a contemporary interpretation of trade secrecy law seems to align with the notion of trade secrets as intellectual property (IP), its historical roots may suggest otherwise.

In the U.S., for instance, trade secrets law originally resembled torts law rather than the "right of property in the idea."[457] Over time, it became part of unfair competition law, labor law, and contracts law.[458] The assertion of trade secrets rights was thus historically more reactive, requiring a violation to occur, such as information theft or breach of an agreement. In contrast, today's perspective on trade secrets as IP offers trade secret owners proactive rights assertion opportunities. Viewing trade secrets as IP, rather than as torts, emphasizes secrecy over wrongful conduct.[459] The owners are deemed to have property-like exclusive rights that can be exercised against any party, including data subjects. As trade secrets as IP has allowed U.S. businesses to keep any information away from public scrutiny as they wish, Kapczynski highlights the importance of adhering to the historical roots of trade secrets law to limit corporate power over information.[460]

Furthermore, data protection authorities are advised to examine the subject matters protected by trade secrecy laws in their respective jurisdictions. While it is commonly understood that any information that is *secret*, *valuable*, and *kept secret* can be protected as a trade secret, there may exist certain hidden definitions or details under the law that the authorities may highlight to prevent potential abuse claims.

For instance, in the U.S., the *secrecy* requirement used to entail only the information that is not readily ascertainable, considering the level

[457] Kapczynski, A. (2022). The public history of trade secrets. *University of California Davis Law Review, 55*, 1387.

[458] ibid., 1384-1386.

[459] ibid., 1394.

[460] ibid., 1440 – 1442.

of effort and difficulty required to obtain such information.[461] As such, some inferences that are more obvious than others (e.g., gender from a photo) may not qualify for trade secret protection to begin with. Furthermore, the *valuable* requirement of U.S. trade secrets law should be understood as protecting only secret information that gains independent economic value from not being generally known to others.[462] In other words, there must be a clear connection between the information's value and its secrecy.

Consequently, for inferences to be protected as trade secrets, businesses should demonstrate that when such inferences become known, their commercial value will diminish. This can be quite challenging for businesses to prove. In addition, some suggest that American trade secrets do not last indefinitely as long as they are *kept secret*; they must continuously provide commercial value.[463] If not, trade secrets will be abandoned. The concept of trade secret abandonment could potentially impose a burden on companies, requiring them to demonstrate their records of using inferences and related commercial gains when denying DSARs, which may also lead to useful disclosures.

## 10.8 Require data controllers to document access requests reasons in detail.

In addition, data protection authorities are advised to require data controllers to document how they handle DSARs in detail. This Article suggests that, when rejecting a DSAR related to inferences, the data controllers must bear the burden of clarifying how complying with the request would harm their trade secrets rights. The clarification on the matter can be part of the Record of Data Processing Activities (ROPA)—or any similar documentation requirement in their jurisdiction—which must be made available for the authority for compliance auditing. Note that the GDPR at Article 30 does not require data controllers to document DSARs refusal grounds in its ROPA. However, this practice is at least suggested by the data

[461] Dole, R. F. (2016) The contours of American trade secret law: What is and what isn't protectable as a trade secret. *SMU Science and Technology Law Review, 19*, 99 (discussing the Restatement (First) of Torts).

[462] Johnson E. (2010). Trade secret subject matter. *Hamline Law Review, 33*, 556.

[463] Lemley, M. A. & Hrdy, C. A. (2021) Abandoning trade secrets. *Stanford Law Review, 74*, 4.

protection authority in Singapore,[464] and this Article proposes that other jurisdictions adopt it to address DSAR denial abuses.

## 10.9 Conclusion and Suggestions for Further Studies

As AI and data analytics tools continue to advance and the cost of collection and processing of personal data becomes cheaper, the practice of inference generation is gaining prevalence. To counter potential misuse of trade secrets law for concealing inferences from data subjects, this Article presents a framework for data protection authorities in any jurisdictions to adopt. This framework prompts the authorities to: 1) critically assess the nature and scope of trade secrets law in their jurisdiction; and 2) require data controllers to document the reasons for denying DSARs pertaining inferences.

Further, the subsequent remedies for data subjects upon becoming aware of inferences about themselves is a crucial aspect for consideration, although outside scope of this Article. Scholars may explore whether the right to correction, for example, can effectively address concerns arising from inaccurate inferences, given their probabilistic nature rather than being outright facts.[465] The right of deletion, for another instance, should also be assessed whether it is practical when applying to inferences. As the input data and the algorithmic model used to generate these inferences remain unaltered, there exists a possibility that that the "algorithmic shadow" could regenerate the same inferences even after deletion.[466]

While these issues remain subjects for further studies, it is of importance that data protection authorities take a proactive step of challenging the application of trade secrets law. This serves as a starting point to ensuring transparency over AI-generated inferences.

[464] *See* Personal Data Protection Committee Singapore. (June 9, 2016). Guide to handling access requests, 9. Retrieved from <https://www.pdpc.gov.sg/-/media/Files/PDPC/PDF-Files/Other-Guides/guide-to-handling-access-requests-v1-0-(090616).pdf> ("Your organization should keep a record of all access requests received and processed, documenting clearly whether the requested access was provided or rejected. Proper documentation may help your organization in the event of a dispute or an application to the PDPC for a review").

[465] Article 29 Data Protection Working Party (2016). Guidelines on automated individual decision-making and profiling, 18 (suggesting false inferences are not necessarily inaccurate if they are statistically correct).

[466] Li, T. C. (2022). Algorithmic destruction, *Southern Methodist University Law Review, 75*, 482.

# 11 Bridging traditional corporate governance and technology: the "AI by Corporate Design" framework to Computational Corporate Governance models

Giuseppe Cicu, University of Torino, Galgano Law Firm

## Abstract

Emerging technologies such as artificial intelligence and big data are rapidly reshaping social, political, and economic environments. The ongoing technological revolution is also influencing the organization and operation of businesses, leading toward the rise of new forms of corporate governance in which AI is adopted in several managerial functions. In spite of the related benefits, the uncritical implementation of AI technologies within a corporate structure ill-suited to accommodate it entails inherent risks, such as lack of transparency and accountability. To prevent a risky adoption of AI within the corporation's structure, the paper proposes a framework called "AI by Corporate Design", aimed at implementing the AI solutions in corporations through an oriented reengineering of the set of mechanisms and processes under which a company operates, upholding respect for the relevant ethical, legal, and algorithmic instances. The proposed framework is grounded on a combination of corporate governance rules, business process management (BPM) techniques, well-known legal disposition (such as the so called "privacy by design"), recommendations and interim regulation regarding the responsible use of AI issued by regulators and governments. While the framework presented in this paper may be voluntary followed by all of the corporations, it appears de iure condendo desirable to locale such "corporate-technological" dimension into the realm of director's duties, as is provided in the Italian legislation under the art. 2086, co. 2, Code Civil, with reference to the directors' duty to establish organizational, administrative, and accounting structures appropriate to the size and nature of the enterprise.

## 11.1 Introduction

Emergent technologies such as artificial intelligence (AI) and big data, characterized by a simultaneous and breathtaking improvements over recent decades, are becoming pervasive in the social, political and economic environments.

The ongoing technological and digital revolution is also affecting the organization and operation of businesses, as well as the processes through which a corporation is administered and managed.

For instance, in 2014, Deep Knowledge Ventures, a Hong Kong venture capital firm, announced to media that it appointed an artificial intelligence program – named *Vital* – capable of making investment recommendations to its board of directors.[467] Even if legally speaking Vital has not acquired the status of corporate director under the corporate laws of Hong Kong, nor the equality in voting on all the financial decision made by the company, it is already known as the "*world first intelligence company director*".[468]

On April 2016, the first blockchain venture capital fund was established in the form of a "decentralized autonomous organization" ("DAO"). Such kind of organization is characterized by decentralization, automatic transaction governance, transparency, and token-based membership. They are designed to challenge traditional hierarchical management structures and the typical division between ownership and control. Indeed, the main goals of the project were to create an organization in which participants would have maintained direct real-time control of contributed funds through governance rules formalized, automated and enforced using smart contract technology.[469] In recent years, DAO's features have been (fully or partially) implemented into the structure

[467] *Algorithm appointed board director,* BBC (May 16, 2014), available at <https://www.bbc.com/news/technology-27426942>. For more details, cfr. Burridge N. (2017), *Artificial intelligence gets a seat in boardroom*, Nikkei Asian Review, available at <https://asia.nikkei.com/Business/Artificial-intelligence-gets-a-seat-in-the-boardroom>.

[468] Zolfagharifard E. (2014), *Would you take order from a Robot? An artificial intelligence becomes the world's first company director*, Daily Mail, available at <https://www.dailymail.co.uk/sciencetech/article-2632920/Would-orders-ROBOT-Artificialintelligence-world-s-company-director-Japan.html>. See also Moslein F., *Robots in the Boardroom: Artificial Intelligence and Corporate Law*, Barfield, W. & Pagallo, U. (eds), Research Handbook on the Law of Artificial Intelligence, pp. 649-670.

[469] Jentzsch C. (2016), *Decentralized Autonomous Organization to Automate Governance*, available at <https://lawofthelevel.lexblogplatformthree.com/wp-content/uploads/sites/187/2017/07/WhitePaper-1.pdf>.

of several corporations, especially in markets like cryptocurrency, DeFi and NFT.[470] Moreover, its characteristics have been shaped by the improvements in AI technologies, whose superior performance in data collection, valorisation and processing has already affected most corporate governance operations: from the monitoring function, through strategy setting and decision-making, to risk management and compliance.[471] Thus, a new type of organization characterized by the combination of AI and blockchain technologies has been theorized: the "AI DAO". In this regard, it has been posited that the decentralized infrastructure of a DAO, along with the use of blockchain combined with AI for decision-making, could led to a self-driving decentralized corporation that operates without management.[472]

Recently, dated back to August 2022, is the announcement of NetDragon Websoft Holdings Limited (a Chinese company that develops and operates massively multiplayer online games in addition to making mobile applications), regarding the appointment of an AI-powered virtual humanoid robot (the so called "Ms. Tang Yu"), as the rotating CEO of its flagship subsidiary, Fujian NetDragon Websoft Co., Ltd. The company declared that "Ms. Tang Yu" serves to streamline process flow, to enhance quality of work tasks, and to improve speed of execution, by helping as a real-time data hub and analytical tool to support rational decision-making in daily operations, as well as to enable a more effective risk management system.[473]

Although there is no unanimous opinion among commentors to what extent does technologies breakthroughs will change corporate governance, the aforementioned examples show the relentless interpenetration between the corporate and algorithmic-computational field and the necessity of re-examining the existing regulatory models

[470] DuPont Q. (2017), *Experiments in algorithmic governance: A history and ethnography of "The DAO," a failed decentralized autonomous organization*, in *Bitcoin and Beyond: Cryptocurrencies, Blockchains and Global Governance*, Routledge. See also, Simonite T. (2016), *The "Autonomous Corporation" Called the DAO Is Not a Good Way to Spend $130 Million*. MIT Technology Review, available at <https://www.technologyreview.com/2016/05/17/160160/the-autonomous-corporation-called-the-dao-is-not-a-good-way-to-spend-130-million/>.

[471] Armour J., Enriques L., et al. (2018), *Putting Technology to Good Use for Society: The Role of Corporate, Competition and Tax Law*, 6 J. BRIT. ACAD, 285.

[472] Bayern S. (2017), *Company Law and Autonomous Systems*: *A Blueprint for Lawyers, Entrepreneurs, and Regulators*, Hastings SCI. & TECH. L.J.,135.

[473] <https://www.prnewswire.com/news-releases/netdragon-appoints-its-first-virtual-ceo-301613062.html>.

of corporate governance and the set of mechanisms and processes under which a company operates, in light of the radical changes and risks arisen by the extensive use of AI technologies.[474]

To do so, I believe that it is necessary adopt an interdisciplinary perspective that includes law, economics, and technological knowledge, deeply combined with an ethical and human-centered dimension. Indeed, echoing the words of Mark Coeckelbergh: "*The technology is always also social and human: AI is not only about technology, but also about what humans do with it, how they use it, how they perceive and experience it, and how they embed it in wider social-technical environments*".[475]

Under these premises, the paper is structured as follows. Section 2 delves into the impact of AI and big data on corporate governance and introduce the emergence of a new kind of organizational governance structure, termed the "computational corporate governance" model. Section 3 introduces the "AI by Corporate Design" as a new framework aimed at integrating AI and big data technologies into business processes and the organizational governance structure, while simultaneously addressing and mitigating risks from privacy and transparency to corporate accountability . Section 4 advocates for the acknowledgment of the technological dimension as an additional fourth pillar of corporations, alongside the organizational, administrative, and accounting structures. Section 5 concludes.

## 11.2 The impact of AI and big data on corporate governance: toward a computational corporate governance model?

Among scholars, it is commonly acknowledged that the superior performance in data collecting, valorisation and processing of AI – including machine learning and deep learning algorithm – when combined with big data technologies, will significantly affect all

[474] Fenwick M., Vermeulen E.P.M. (2018), *Technology and Corporate Governance: Blockchain, Crypto, and Artificial Intelligence*, ECGI Working Paper Series in Law. Zagar M.T. (2018), *A New Chapter for ICONOMI: Transformation of Corporate Governance and Issuance of Equity Tokens*, *Medium*, September 2018, available at <https://medium.com/s/the-bigdisruption/the-big-disruption-36fbed0268cf>. For an in-depth analysis of the different positions in scholarship, see Enriques L. and Zetzsche D.A., *Corporate Technologies and the Tech Nirvana Fallacy*, Hastings Law Journal, 2019.

[475] Coeckelbergh M. (2020), *AI Ethics*, The MIT Press Essential Knowledge Series, 79.

corporate governance operations: from monitoring function and strategy setting to decision-making to and compliance.

However, the measure of how AI and big data will affect corporate governance remains debated. In this regard, it is possible to refer to two macro-opinions.

In one hand, there are academics who states that such technologies will change existing corporate governance paradigm and will overcome well-known corporate governance problems (such as the agency problem[476]). It can also be linked to this view the opinion of who assumes that in the future AI algorithm will not only assist directors and managers, but will also replace them in their decision-making, serving itself as a sole board member. To the purpose of this paper, we can call them "Tech Proponent".

On the other hand, there's a more moderate opinion under which AI and big data technologies will improve governance procedures and practices without making the corporate boards and managers' functions obsolete. Them, while supporting the fact that corporate boards will be supported and improved by the implementation of these technology, contest the adversarial assumption under which such technologies will replace board or significantly change the core function of board and management.

It appears that the core difference between the above-mentioned opinions lies in the (more optimistic or pessimistic) view on the application and future potential of AI and big data.

Specifically, Tech Proponents assume that within a business organization AI can already (i) supports corporate functions and improves board-decision making (the so called "assisted AI")[477] and/or (ii) provides "*support in more complex problem solving and decision-*

[476] Agency theory predicts that the divergences of interests between managers and shareholders could lead to "agency problem", that is, managers engage in activities for their own self-interest rather than the benefits of the shareholders. The costs experienced by the principal to limit this misalignment of interests are known as "*agency costs*", defined as the sum of the monitoring expenditures by the principal, the bonding expenditures by the agent, and the residual loss. See Michael C. Jensen C.M. and Mecklin H.W. (1976), *Theory of the firm: managerial behaviour, agency costs and ownership structure*, Journal of Financial Economics, 3, 305.

[477] Examples of commonly used AI systems of this nature are Apple's Siri and its Android rival, Google Assistant. See Wadhwa V. and Salkever A. (2017),*The Driver In The Driverless Car*, 38. Rao A. (2016), *Ai Everywhere & Nowhere* Part 3 – AI is AAAI (Assisted-Augmented- Autonomous Intelligence), Pwc Next In Tech, available At <<http://Usblogs.Pwc.Com/ Emerging-Technology/ Ai-Everywhere-Nowhere-Part-3-Ai-Is-Aaai-Assisted-Augmentedautonomous- Intelligence>.

*making situations by asking and answering questions as well as building scenarios and simulations*." (the so called "augmented AI").[478]

In this regard, it has been said that assisted and augmented AI will soon be able to replace corporate boards in doing the administrative tasks, by doing the relevant activities faster, better, and at a lower cost. To summarize this scholarly perspective, they assert that in the near future "*AI will put an end to administrative management work*".[479] According to this view, the implementation of AI in the boardroom will take away time-consuming administrative tasks, such as monitoring and reporting, from the components of the board, allowing them to dedicate more on judgements-based activities, like setting the company's strategy.

Tech Proponents also assume that AI will enhance the decision-making process of the board of directors.[480] For instance, the AI will play an important role in the evaluation of a merger or an acquisition, by instantly analysing the amount of data at a firm's disposal, creating accurate reports and suggesting the best decision to take, while considering the relevant regulation.

Finally, from an economic perspective, the functions performed by AI will significantly lower the cost of administrative and predictions tasks.[481] Considering the scenario in which AI would be put in place to assist or increase the human performance, the component's board involvement shall still be present, specifically to make the final decision on the basis of the available assessments carried out by the AI algorithms.

[478] Examples for the category of advisory or augmented AI include IBM's Watson platform. (Forrest C. (2015), IBM Watson: What are companies using it for?, ZDNET available at ,https://www.zdnet.com/article/ibm-watson-what-are-companies-using-it-for. Kolbjørnsrud V. Et Al. (2016), *The Promise Of Artificial Intelligence: Redefining Management In The Workforce Of The Future* 6, available at <https:// www.accenture.com/us-en/insight-promise-artificial-intelligence>.

[479] Kolbjørnsrud V. Et Al. (2016), *The Promise Of Artificial Intelligence: Redefining Management In The Workforce Of The Future*, at 3. The study mentions tasks such as note taking, scheduling, reporting, maintaining scorecards, managing shift schedules, and generating investor statements and management reports as specific examples of AI-led administrative work. Id. at 4, 11, 17.

[480] Beck M., Libert B. and Bonchek M. (2017), *AI in the Boardroom: The Next Realm of Corporate Governance,* MIT Sloan Management Review, available at: <https://sloanreview.mit.edu/article/ai-in-the-boardroom-the-next- realm-of-corporate-governance/.>

[481] Agrawal A., Gans J. and Goldfarb A. (2016). *The Simple Economics of Machine Intelligence*. Harvard Business Review, available at: <https://hbr.org/2016/11/the-simple-economics-of-machine-intelligence>.

Looking further ahead, Tech Proponents envision a future where a technological board governed by AI algorithms – termed "Algo-Board" – will replace traditional board members. Unlike a conventional board of directors led by humans, the Algo-Board would be governed by algorithms. This system would process vast amounts of information, evaluate strategic options in real-time, and make data-driven decisions aligned with the company's objectives and values. There are three assumptions that ground this view: (i) the development of the so called "general artificial intelligence", an AI able to understand or learn any intellectual task that a human being can[482]; (ii) AI solutions will be able to perform both administrative and judgment tasks better than humans[483]; (iii) humans will become less fit to serve as board members than machines, or will be less willing to do so.[484]

The commentors of the more moderate opinion on the impact of AI and big data within corporate governance – despite the fact that they do not contest the beneficial effect of the application of these technologies – assume that, whether the human members of the board were to be replaced by AI algorithm, decision would not be better than they already are from the shareholder view.

Therefore, they predict a scenario in which such technologies have a more limited role in the boards, such as informing board members about option and opportunities, without replacing them neither in monitoring nor in mediating function.[485]

Indeed, in light of the increasing complexities and risks inherent in AI, the proposition of an entirely AI-driven corporation – despite debates on its viability – seems not desirable. Alternatively, it seems auspicious the rise of a new form of organizational corporate governance structure, termed "computational corporate governance" model, in which the main features of AI technologies (e.g., predictive analysis

[482] Hal H. (2019). *DeepMind and Google: the battle to control artificial intelligence*; Henry S., Karina V., Matthew C. and Marta H. (2019), *The limits of machine intelligence: Despite progress in machine intelligence, artificial general intelligence is still a major challenge*". EMBO Reports, 20.

[483] UCL Working Paper Series *Corporate Management in the Age of AI*, 30 (No.3/2019).

[484] See Fenwick M. and Vermeulen E.P.M. (2019), *Technology and Corporate Governance: Blockchain, Crypto, and Artificial Intelligence*, 48 TEX. J. BUS. L. 1, 2.

[485] Enriques L. and Dirk A. Zetzsche (2019), *Corporate Technologies and the Tech Nirvana Fallacy*, Hastings Law Journal p. 74.

systems, automated decision-making, natural language processing, and so on) are ethically integrated into the board's functions/structure and corporate's processes, without sacrificing the role of the human being at the altar of technology. However, to achieve this outcome, there's a pressing need to rethink the standard rules of corporate governance in alignment with technological advancements.

## 11.3 "AI by Corporate Design": a proposed framework to govern the corporate governance transition toward the new technological paradigm.

The increasing impetus toward incorporating AI and big data systems within business processes and governance structures highlighted the countless risks associated with their indiscriminate adoption. These risks, if unaddressed, may propagate throughout the organization structure, potentially compromising both the corporate functions and the stakeholders' rights. Such risks stem from the extremely complex nature of the AI, encompassing issues related, inter alia, to privacy, ethics, transparency, and accountability matters.

Consequently, as stated by Floridi "*the real challenge is no longer digital innovation, but the governance of the digital*", which he describes as "*the practice of establishing and implementing policies, procedures, and standards for the proper development, use and management of the infosphere*".[486]

However, in the realm of business and corporate governance, it is crucial to acknowledge that corporations must follow their own rules, policies, procedures, and standards in pursuing their entrepreneurial objectives. Hence, the mere juxtaposing of AI-related procedures alongside established corporate regulation it is not sufficient to ensure the secure, effective, and efficient integration of such technology into the business structure (just as the simple purchase and utilization of highly performant software does not constitute enterprise digitization).

In light of the above, this paper proposes a framework referred to as "AI by Corporate Design".

[486] Floridi L., *Soft Ethics and the Governance of the Digital*, 2018, 3.

This framework is designed to assist corporations in effectively integrating and managing AI and big data technologies, with a particular focus on (i) identifying, evaluating, preventing, or mitigating risk factors related to the use of AI systems; (ii) optimizing the benefits derived from AI tools; (iii) supporting the members of corporate governance structure in executing.

its strategic, administrative, and supervisory roles; (iv) promoting a sustainable and human-centric approach to of AI utilization.

Drawing inspiration, among others, from the "privacy by design" concept and its guiding principles,[487] this framework (i) advocates for proactive measures over reactive ones; (ii) emphasizes the importance of preliminary impact analysis for AI and big data technologies on stakeholders' rights; (iii) embeds AI ethics into the design of corporate processes and corporate governance structure; (iv) provides clear, phase-specific transparency and accountability policies throughout the AI lifecycle stages; (v) fosters a user-centric approach to AI and big data utilization in relation to corporate operations.

The "AI by Corporate Design" framework draws considerable insight from other AI-focused frameworks, such as the "Ethics Guidelines for Trustworthy AI" by the European Commission[488] and the "Fairness, Accountability, and Transparency in Machine Learning" (FAT/ML) guidelines,[489] and the "Accountable Algorithms" principles.[490] However, unlike these frameworks, "AI by Corporate Design" framework distinguishes itself by adopting a holistic approach specifically dedicated to corporations and enterprises that simultaneously encompasses the three pivotal dimensions of 1)

[487] *Privacy by Design* is a concept developed by Dr. Ann Cavoukian in the '90s to address the systemic effects of Information and Communication Technologies and of large-scale data systems. This concept advances "*that future of privacy cannot be assured solely by compliance with regulatory frameworks; rather, privacy assurance must ideally become an organization's default mode of operation*". As Cavoukian stated, the objectives of *Privacy by Design* may be accomplished by practicing the following seven foundational principles, extensible to the "AI by Corporate Governance" framework: 1) Proactive not reactive; Preventive not Remedial; 2) Privacy as the Default Setting; 3) Privacy Embedded into Design; 4) Full Functionality – Positive-Sum, not Zero-Sum; 5) End-to-End Security – Full Lifecycle Protection; 6) Visibility and Transparency – Keep it Open; 7) Respect for User Privacy – Keep it User-Centric. See Ann Cavoukian, 2011, available at <https://www.sfu.ca/~palys/Cavoukian-2011-PrivacyByDesign-7FoundationalPrinciples.pdf>.

[488] Available at <https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai>.

[489] Available at <https://www.fatml.org/resources/principles-and-best-practices>.

[490] Available at <https://oecd.ai/en/catalogue/tools/principles-for-accountable-algorithms-and-a-social-impact-statement-for-algorithms>.

corporate governance and business processes; 2) legal regulation, emphasizing both privacy (and its inherent need for transparency) as well as corporate law (with the relevant focus on accountability), and 3) technological dimension.

Specifically, the "AI by Corporate Design" framework aims at strategically and ethically integrating AI and big data technologies within two core layers of corporate structures: "corporate processes" and "corporate governance".

## 11.4 An AÍ Business Process Management (BPM)

With respect to corporate processes, the cornerstone of this framework is the "Business Process Management" (BPM). BPM is a business field that encompasses the identification, visualization, design, execution, monitoring, and optimization of business processes. Traditionally, BPM involves a cycle comprising the following stages: 1) process strategic planning; 2) process design; 3) process visualization; 4) process implementing; 5) process evaluation and monitoring; 6) process optimization.[491]

Within the context of the "AI by Corporate Design" framework, the six phases of BPM are aligned with the phases of an AI lifecycle model inspired by the Cross-Industry standard process for Data Mining (CRISP-DM).[492] Consequently, to facilitate the integration of AI into corporate processes following the BPM stages, an AI lifecycle model is proposed, comprising the following six phases: 1) AI problem and goal definition; 2) AI data collection/pre-processing and AI selection and design; 3) AI training, validation, and testing; 4) AI model deployment; 5) AI evaluation and monitoring 6) AI improvement.

As a result, within the "AI by Corporate Design" framework, stages of BPM and AI lifecycle converge, aiming to construct a technological corporate architecture adept at managing the multifaceted

[491] ABPMP International (2019), BPM CBOK, *Guide to Business Process Management Common Body of Knowledge*, 29. See also, Szelagowski M. (2018), *Evolution of the BPM Lifecycle*, in Communication Papers of the Federated Conference on Computer Science and Information Systems, p. 205.

[492] CRISP-DM aims to offer a comprehensive framework for executing any project employing scientific methods to extract value from data, including Machine Learning. CRISP-DM divides a project into six phases: 1. business understanding; 2) data understanding; 3) data preparation; 4) modeling; 5) evaluation; 6) deployment. See Haakman, M., Cruz, L., Huijgens, H. et al. (2021), *AI lifecycle models need to be revised*, Empirical Software Engineering, 26, 95.

considerations related to AI, while concurrently preserving the intrinsic characteristic of a corporation. To achieve this, the following roadmap can be pursued:

### 11.4.1 BPM: Process Strategic Planning & AI Lifecycle Problem and Goal Definition

BPM Process Strategic Planning comprises the subsequent sub-stages:

- Process-driven strategy development: this phase fosters an understanding of organizational strategies and goals designed to pursue the corporation's purpose.
- Stakeholder Engagement: actively involving key stakeholders provide insights into potential areas for improvement.
- Key Performance Indicators (KPIs) Identification: during this phase, (KPIs) relevant to business processes and AI technologies are identified. These KPIs are instrumental not only in gauging the efficacy of the processes but also in assessing the seamless integration and performance of AI within the organizational structure. The utilization of KPIs provides a quantitative measure to evaluate the congruence between processes reengineering, AI implementation and the overarching corporate objectives.
- Goal Definition: this phase involves establishing clear and quantifiable objectives that the business processes aim to accomplish following reengineering and subsequent AI integration. By defining these goals, corporations can set a benchmark for success and ensure that both the redesigned processes and the AI systems work in tandem to achieve the desired outcomes.

AI Lifecycle Problem and Goal Definition, encompasses the following sub-stages:

- Feasibility Analysis: on the basis of the BPM's first phase outcomes, this stage conducts an initial examination to determine whether AI solutions can address their designates objectives and the identified inefficiencies in processes, as well as mitigate potential vulnerabilities concerning stakeholders' rights. For corporation that have already adopted AI technologies, an assessment is conducted to determine the technologies' alignment with the safeguarding of stakeholders' rights.

Scope of AI Application: during this phase the potential scope of AI solutions is assessed, laying groundwork for choosing appropriate algorithm models.

- Goal Specification: this stage focuses on articulating clear AI objectives to ensure that the proposed AI interventions align with the broader business goals and strategic direction identified during the initial BPM's phase.
- Data Mapping: This phase is instrumental in understanding and visualizing the intricate flow of data within an organization. By systematically identifying the origins, touchpoints, and destinations of data, the process ensures that the entirety of the data's lifecycle is charted. This comprehensive mapping not only facilitates a clearer comprehension of how data traverses through various systems and processes but also identifies potential bottlenecks, redundancies, or gaps. In the context of AI integration, such a precise map becomes indispensable, as it ensures that the AI systems have access to accurate and relevant data, while also highlighting areas where data integrity and privacy may be at risk.

### 11.4.2 BPM Process Design & AI Lifecycle Data Collection/ Pre-processing and AI Selection and Design

BPM Process Design, encompasses the following sub-stages:

- Gap Analysis: this phase involves an in-depth review and analysis of existing processes to identifies inefficiencies, redundancies, or bottlenecks. The objective is to gather a comprehensive understanding of areas that might be hindering optimal performance or productivity.
- Processes Reengineering: the goal of this phase is redesign and refine the processes, leading to the formulation of streamlined, efficient workflows that align with organizational objectives.

AI Lifecycle Data Collection/Pre-processing and AI Selection and Design, encompasses the following sub-stages:

- Data Collection/ Pre-processing: drawing from the findings this phase is dedicated to select and gathering pertinent data for AI solutions. Emphasis is placed on data quality and volume, ensuring its relevance and suitability for subsequent stages. It's essential that data is acquired and processed ensuring the respect of stakeholder's rights while upholding transparency and accountability, in alignment with best practice and regulatory guidelines. For corporation that have already adopted AI technologies, an assessment is conducted to verify the data acquisition's compliance with stakeholders' rights. Moreover, at this stage, data is cleansed, normalized, and prepared for model training. Specific techniques, such "data synthetic, might be chosen to enhance the protection of stakeholder's rights.

- AI Selection and Design: depending on the specific nature of the needing and the available data, appropriate algorithms are chosen to best address the identified corporate requirements. Furthermore, frameworks and architectures for the selected AI algorithms are designed, priming them for the training phase.

### 11.4.3 BPM Process Visualization & AI Lifecycle Model Training, Validation and Testing

BPM Process Visualization, encompasses the following sub-stages:

- Future State Visioning: this step involves the conceptualization of potential process states following both reengineering and the incorporation/analysis of AI solutions.
- Simulation: Using advanced tools, this phase simulates the outcomes of prospective process modifications, thereby forecasting the impacts of AI integration both on business operations and on the safeguarding stakeholders' rights.

AI Lifecycle Model Training, Validation and Testing, encompasses the following sub-stages:

- Training: In this phase, the models undergo comprehensive training using the preprocessed data. The model continually refines its internal parameters and weights to reduce prediction errors and enhance performance, ensuring alignment with the foundational corporate objectives and stakeholders' rights.
- Validation: During this phase, a distinct set of data, known as the validation set, is utilized. This data has not been a part of the model's training. The primary objective here is to refine the model parameters, ensuring the model does not overfit to the training data, thus ensuring its generalizability.

Testing: Once the model has been trained and validated, it undergoes testing using an entirely new dataset that it hasn't encountered before, referred to as the test set. This stage critically evaluates the model's predictive capabilities in real-world-like scenarios, gauging its readiness for deployment.

### 11.4.4 BPM Process Implementing & AI Lifecycle Model Deployment

BPM Process Implementing, encompasses the following sub-stages:

- Implementation: reengineered processes are operationalized in the corporate structure, encapsulating the outcomes of previous stages. Throughout this phase, an effort is made to ensure that the redefined processes are in alignment with and respectful of stakeholders' rights

AI Lifecycle Deployment, encompasses the following sub-stages:

- Deployment: AI models are meticulously incorporated into the restructured business processes. This integration aims to maximize the potential of AI, ensuring congruence with the operational framework and effectively addressing the previously identified corporate requirements. Throughout the integration, consistent attention is given to ensure that the deployment of AI solutions remains compliant with and protective of stakeholder rights.

### 11.4.5 BPM Process Evaluation and Monitoring & AI Lifecycle Model Evaluation and Monitoring

BPM Process Evaluation and Monitoring, encompasses the following sub-stages:

- Real-time Monitoring: advanced tools and specified metrics are deployed to continuously observe and record the performance of the reengineered processes in real-time. This monitoring ensures that the processes remain adaptive and responsive to any discrepancies, while also emphasizing the importance of safeguarding stakeholder rights.
- Performance Analysis: drawing upon the previously defined KPIs, the performance of the processes is periodically scrutinized. This analysis provides a structured feedback loop to assess the

effectiveness of the implemented changes, and their alignment with stakeholders' rights.

AI Lifecycle Model Evaluation and Monitoring, encompasses the following sub-stages:

- Model Performance Monitoring: continuous tracking tools evaluate the AI models' performances post-deployment, ensuring their accuracy, efficiency, and compliance with stakeholder rights.
- AI Impact Analysis: periodically, the influence and implications of AI solutions on both the operational and stakeholder dimensions are assessed. This review ensures that the AI implementations remain transparent, ethical, and in line with the broader corporate objectives while respecting stakeholders' rights.

### 11.4.6 BPM Process Optimization & AI Lifecycle Improvement

BPM Process Optimization, encompasses the following sub-stages:

- Feedback Loops: instituted continuous improvement mechanisms capture feedback from various process touchpoints, ensuring iterative refinement of processes. Such loops emphasize not only on operational efficiency but also on ensuring that processes are consistently aligned with the safeguarding of stakeholder rights and interests.

AI Lifecycle Improvement, encompasses the following sub-stages:

- Model Optimization: Informed by real-world performance data, iterative adjustments and refinements are made to AI models. These adjustments aim to bolster accuracy, reduce latency, and enhance other pivotal performance metrics, all while ensuring that the models' functions remain transparent, ethical, and in compliance with stakeholder rights and expectations.
- Trough the adoption of the "AI by Corporate Design" framework, the synergies between BPM and AI empowers corporations to achieve a balance between corporation structural efficiency, technological adoption, and the safeguarding of stakeholders' rights, especially with regards AI transparency and accountability instances.

Specifically, as following elucidated:

**(i)** BPM promotes the implementation of standardized practices and processes to ensure consistency, efficiency, and clarity in corporate operations. When AI and big data systems are developed and deployed within this structure, there is a consistent method, policies, and visualization to their operation. This predictability not only ensures that AI functions are transparent but also increase trust among stakeholders since operations and decisions made by AI can be understood and reviewed methodically within the established processes and policies.

**(ii)** With BPM, its processes, data mapping and their visualization in place, any decision made by the AI can be traced back to its source data and logic. This traceability ensures transparency and accountability as stakeholders can understand the rationale behind AI decisions.

**(iii)** As corporations methodologically assess and manage AI risks within the corporate structure, any potential pitfalls or biases can be identified and rectified in advance. This proactive approach safeguards against unintended consequences and, by extension, promotes transparency in how AI models are developed, implemented, utilized, and improved.

**(iv)** By placing a primal attention on stakeholders' rights, the model inherently stresses the ethical use of AI. Ethical AI is transparent and accountable by design, prioritizing fairness, and explicability.

**(v)** The synergy between BPM and AI facilitates an iterative feedback loop. Should an AI system operate in an unexpected or undesirable manner, this feedback mechanism ensures prompt resolution, thereby holding corporations accountable for any discrepancies.

**(vi)** With the establishment of resilient and robust governance structures, clear delineations of responsibilities regarding AI deployment, implementation, and utilization emerge. When roles and expectations are clearly defined, accountability is inherently heightened. This ensures that any inconsistencies in AI transparency and accountability are promptly identified and rectified.

**(vii)** Prioritizing stakeholders' rights shifts their role from passive observers to active contributors. Their active involvement in the process ensures that AI systems are conceptualized, designed, and iteratively refined in alignment with their heightened expectations for transparency and accountability.

In essence, the harmonization of BPM's systematic rigor and AI's transformative capabilities creates a symbiotic relationship. This partnership champions the causes of transparency and accountability, ensuring that AI, while disruptive and innovative, remains ethically grounded, accessible, and understandable to all stakeholders.

On the corporate governance side, considering the multi-layered complexities that AI introduces in the corporate architecture, particularly in decision-making algorithms and predictive analytics, a dedicated committee on algorithmic use becomes paramount. This committee would be tasked with ensuring that AI implementations not only compliance to legal standards and ethical instances but also align with overarching corporate principles.

Thus, the second main part of the "AI by Corporate Design" framework is represented by the creation of an Ethic, Algorithmic, and Legal Committee ("EALC") within the corporate governance structure.

Specifically, the EALC, comprised of an interdisciplinary team of managers, directors, and consultants, bear the responsibility and the accountability for creating and/or updating the organizational structure in line with the BPM and AI lifecycle phases. From its inception, the primary focus of this committee is to assess the impact of AI technologies on main business processes and relevant stakeholders. Moreover, this committee is strategically designed to champion both transparency and accountability. Acting as a "filter" between AI outputs and stakeholders, it embraces the "human in the loop" principle by putting the human judgment and control over the AI decisions. This ensures transparency on the activities carried out by the committee as well as its accountability for the outcome of the AI processes.

Furthermore, adopting this approach has the potential to strengthen the corporation's market standing and enhance customers engagement, due to its emphasis on ethical considerations. Additionally, embracing such a framework can also position the corporation to align more closely with EU and International AI regulations.

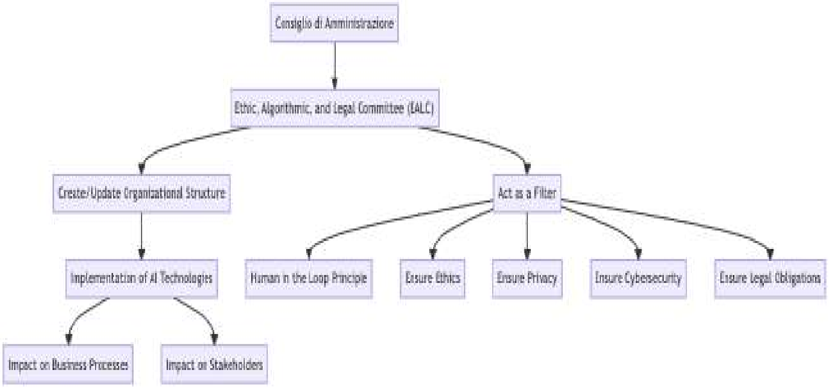

Consiglio di Amministrazione
Ethic, Algorithmic, and Legal Committee (EALC)
Create/Update Organizational Structure
Act as a Filter
Implementation of AI Technologies
Human in the Loop Principle
Ensure Ethics
Ensure Privacy
Insure Cybersecurity
Ensure Legal Obligations
Impact on Business Processes
Impact on Stakeholders

Figure 1. The role of EALC.

## 11.5 De iure condendo: technology as a fourth dimension of an adequate corporation's structure

As AI technologies become more integrated into corporate structures and their governance functions, it becomes imperative to revisit and realign corporate principles in light of the complexities introduced by this transformative technology.

While corporations can proactively embrace best practices and frameworks (like the one proposed in this paper), the profound impact of AI on stakeholder's rights underscores the necessity for legislative reform. Specifically, it is believed that there is a pressing need to revisit the principles of corporate law and corporate governance, integrating the technological dimension as an essential component of contemporary and future-oriented enterprises.

However, given the rapid pace of technological advancements in AI (illustrated, among the others, by the urgency to reevaluate the AI EU ACT in light of developments in the generative AI) it might be unfeasible to foster immediate, specific legislation. Indeed, such kind of legislation could risk becoming obsolete shortly after its enactment, due to the high-dynamic nature of the technology.

Consequently, a more suitable legislative intervention would be to recognize the technological dimension as an intrinsic component within the corporate structure and organization. This is not only to maintain compliance with the law but also to ensure resilience, adaptability, and foresight in a rapidly changing technological landscape.

In the era of digital transformation and given the rising prominence of AI in business operations, the Article 2086, co. 2 of the Italian Civil Code, can emerge as a pivotal reference to integrate the technological dimension as an inherent component of the corporate structure. Specifically, it mandates: “*The entrepreneur, whether operating in corporate or collective form, has a duty to set up an organizational, administrative, and accounting structure appropriate to the nature and size of the enterprise, also with a view to the timely detection of the enterprise’s crisis and the loss of business continuity* (…).

As businesses increasingly adopt digital tools, platforms and advance technologies, the duties and responsibilities of directors must adapt to provide comprehensive oversight of these technological assets within the corporate framework. In this regard, the objective is to ensure that the company remains compliant and resilient in the face of challenges posed by disruptive technology such as AI.

Given the aforementioned consideration, in addition to the three dimensions by Article 2086, co. 2, of the Italian Civil Code – encompassing the organizational, administrative, and accounting structure – it becomes imperative to also acknowledge the technological infrastructure as a fourth dimension within the corporate structure, underscoring the relevant duties and responsibilities of directors.

## 11.6 Conclusion

In this remarkable era, marked by noteworthy technological advancements in AI and big data technologies, corporations face both unprecedented opportunities and challenges. The promise of operational efficacy and strategic advantage through AI integration is counterbalanced by evolving responsibilities related to the corporate field. As new models of corporate governance, such as “DAO” and “Computational Corporate Governance” model, these issues will intensify. The proposed “AI by Corporate Design” framework offers a balanced approach to embedding AI and big data within corporate infrastructures, ensuring compliance with ethical, legal, and technological standards. Given the rapid evolution of technology, there's a pressing need for legislative action that incorporates the technological dimension into corporate frameworks, and consequently into directors' duties and responsibilities.

# 12 Clarifying Military Advantages and Risks of AI Applications via a Scenario

**Liisa Janssens, LLM MA, Scientist Military Operations, Unit Defense Safety and Security, The Dutch Applied Sciences Institute, TNO**[493]

## Abstract

This paper illustrates the necessity to adhere to the tenets of the Rule of Law in order to establish responsible deployment of Artificial Intelligence (AI) in military theatres. Tenets of the Rule of Law are accountability, transparency and contestability; these tenets function together in the mechanisms of the Rule of Law. Examples of existing Rule of Law mechanisms of the legislative, executive and judicial powers are: (re)shaping legislation and formulating (new) policies. AI can be seen as an Emerging Disruptive Technology (EDT) to Rule of Law tenets and mechanisms which disruptiveness needs to be addressed.

In this paper an example of an AI application, which can be deployed in military operations, is investigated via a scenario. In this scenario the risks for upholding the Rule of Law tenets and mechanisms are illustrated via their relation to NATO's principles of responsible use. Furthermore, the possibilities to mitigate these risks are illustrated via examples of how to reshape existing Rule of Law mechanisms. In order to determine what is at stake when deploying AI in military theatres, an interdisciplinary approach is used. This approach brings together law, philosophy and technology (Artificial Intelligence and systems) via a military operational scenario. In the military operational scenario, a Counter Unmanned Aircraft System (C-UAS) is enhanced with an AI application. Via this scenario examples of disruptiveness are given which lead to an illustration to different stakeholders in the separated powers (legislative, executive and judicial) of how pressure on the Rule of Law tenets can be identified. After identification of this pressure,

[493] This paper would not have been possible without the support of the interdisciplinary TNO team of the NATO project 'The Design of AI Applications in Counter Unmanned Aircraft Systems and the Rule of Law' (*report is forthcoming*). The guidance of TNO colleagues Larissa Lobbezoo Msc, Okke Lucassen MA, Laura Middeldorp Msc and Peter Verkoeijen MA, and the forward looking approach by dr. Claudio Palestini and Marie Paulus MA of NATO, were indispensable for the success of this paper.

the next step is answering the question how the tenets of the Rule of Law can be protected by adding *new* requirements to existing Rule of Law mechanisms, i.e. how can *(newly found)* technical requirements enrich *(new)* legislation, interpretations of old legal concepts and/or inform policies? The goal in this research is to showcase, via an interdisciplinary approach in the context of a scenario, how to prevent unintentional harm to the tenets and mechanisms of the Rule of Law. Given that aim is to deploy AI applications in military operations in a responsible way, risks need to be identified and mitigated. This paper informs end-users, policymakers, regulatory authorities, researchers, and industry on the potential added value and the limitations of AI applications, and how to mitigate the possible risks.

## 12.1 Clarifying Military Advantages and Risks of AI Applications via a Scenario

There is a pressing need for research on how to deploy Artificial Intelligence (AI) in a responsible way in military theatres. AI is an example of an Emerging Disruptive Technology (EDT)[494] and contemporary reflections on the nature of law and, especially, its relations to moral reasoning, are challenged by Emerging Disruptive Technologies (EDTs). AI is an example of an EDT that poses challenges to the adherence to core principles of the system of law when AI is used in the context of military operations. The analysis of the introduction and use of AI in a military context, and in particular its risks for adhering to the tenets of the Rule of Law and its mechanisms -without missing opportunities to create military advantages- has become more important than ever before.

To assist in balancing the risks and opportunities of AI applications in military theatres, this research is conducted via an interdisciplinary approach in which the disciplines law, philosophy and (AI) technology are brought together in a Counter Unmanned Aircraft System scenario. In this interdisciplinary approach scenarios are used as a way to seek for *"an integrative level of understanding"*[495] of the

[494] NATO Emerging and disruptive technologies, Retrieved August 2023, from. <https://www.nato.int/cps/en/natohq/topics_184303.htm>.

[495] Austin, W., Park, C., & Goble, E. (2008). From interdisciplinary to transdisciplinary research:.A case study. *Qualitative Health Research*, *18*(4), 557-564.

potential military advantages of AI applications and possible risks to the core principle of the system of law: the Rule of Law.

The conducted interdisciplinary approach in the form of a scenario *(Figure 1)* leads to an illustration for different stakeholders in the separated powers (legislative, executive and judicial) of how pressure on the Rule of Law tenets can be identified. After identification of this pressure, the next step is answering the question how the tenets of the Rule of Law can be protected by adding new requirements to existing Rule of Law mechanisms, i.e. how can *(newly found)* technical requirements enrich *(new)* legislation, interpretations of old legal concepts and/or inform policies?

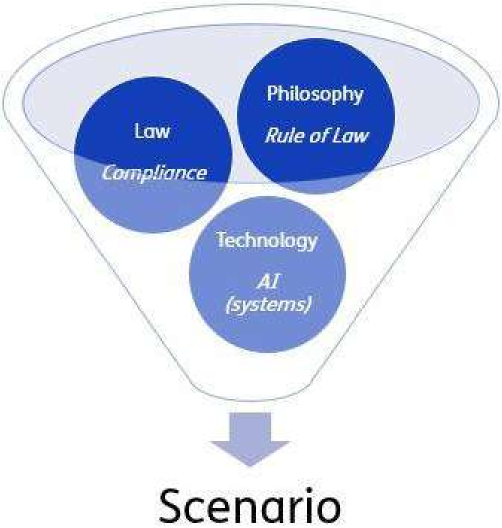


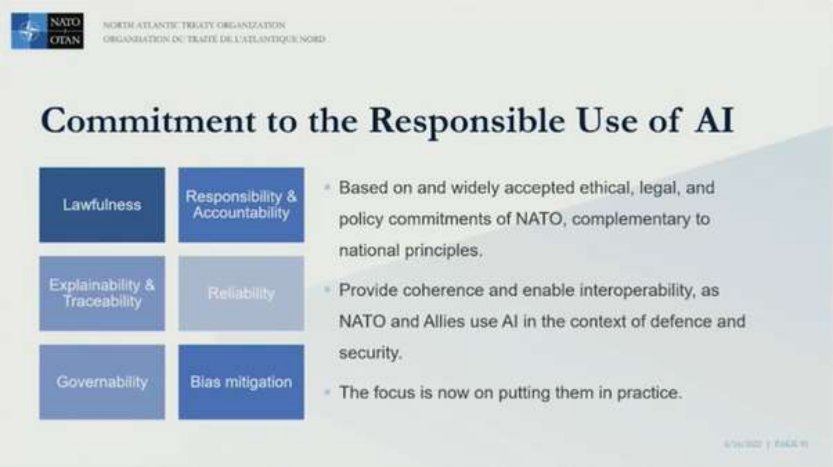
Figure 1: Interdisciplinary approach in the context of scenarios.

This paper first elaborates on the risks to the Rule of Law tenets and mechanisms; these are illustrated via their relation to NATO's principles of responsible use *(Figure 2)*. Furthermore, the possibilities to mitigate these risks are illustrated via examples of how to reshape existing Rule of Law mechanisms. The goal is first to showcase how, via interdisciplinary research in the context of a scenario in which AI is deployed in a military theatre, unintentional harm to the tenets and mechanisms of the Rule of Law can be prevented, and second to demonstrate how this showcase can contribute to NATO's ambition to formulate a toolkit for responsible AI certification standards.

When the aim is to deploy AI applications in military operations in a responsible way, risks need to be identified and mitigated. This paper informs end-users, policymakers, regulatory authorities, researchers, and industry on the potential added value and the limitations of AI applications, and how to mitigate the possible risks.

## 12.2 The Rule of Law in Relation to NATO's Principles of Responsible Use

In this section two NATO principles of responsible use (*lawfulness* and *governability*) are introduced and contextualised via the Rule of Law. What is responsible AI, and what is its relevance with respect to Rule of Law tenets and mechanisms? Responsible AI starts at the acquisition process, or – in case of a newly developed (AI) technology- at the design phase: How can the design in the research & development and/or acquisition processes of AI applications, throughout the whole value chain of partners, be informed with (*new*) requirements in a way that this leads to responsible AI?



NATO OTAN NORTH ATLANTIC TREATY ORGANIZATION ORGANISATION DU TRAITÉ DE L'ATLANTIQUE NORD

Commitment to the Responsible Use of AI

Lawfulness

Responsibility & Accountability

Explainability & Traceability

Reliability

Governability

Bias mitigation

- Based on and widely accepted ethical, legal, and policy commitments of NATO, complementary to national principles.
- Provide coherence and enable interoperability, as NATO and Allies use AI in the context of defence and security.
- The focus is now on putting them in practice.

Figure 2: Illustration of the six principles of responsible use of AI: lawfulness; responsibility & accountability; explainability; reliability; governability; bias mitigation.[496]

[496] NATO presentation (2022) Future conflicts may be won or lost by AI presented by Nikos. Loutas, Head of Data and Artificial Intelligence Body, NATO, in: Tech Informed, Retrieved September 2023, from <https://techinformed.com/nato-future-conflicts-may-be-won-or-lost-by-ai/>.

The Rule of Law[497] is one of the tenets that constitutes democratic societies with checks and balances throughout the whole value chain of partners, where each power *(Figure 3)* has its own protection mechanisms against internal and external power abuse. The United Nations clarifies the Rule of Law mechanisms as follows:

> *"It requires measures to ensure adherence to the principles of supremacy of the law, equality before the law, accountability to the law, fairness in the application of the law, separation of powers, participation in decision-making, legal certainty, avoidance of arbitrariness, and procedural and legal transparency."*[498, 499]

The above-mentioned mechanisms are necessary for a democratic society to function properly and can be seen as a good example of mechanisms which constitute *governability* of AI applications. In the context of these mechanisms the processes before and after deployment of AI applications need to be guided with checks and balances from the legislative, executive and judicial power. For example: when external parties such as manufacturers and/or developers of AI applications cannot be checked by the executive power using *(new)* requirements for what constitute responsible AI, the risk arises that unintentional harm is done to the Rule of Law. And without such *new* requirements, the judicial power might not be equipped, when errors or casualties occur during deployment, to check afterwards if the design, acquisition and/or research and deployment can be qualified as responsible within the rules and regulations of a democratic society. Consequently, this may lead to erosion of NATO's principle of responsible use: *lawfulness*, and consequently also of the tenets of the Rule of Law.

[497] Article 2 of the Treaty on European Union: *"The Union is founded on the values of respect for human dignity, freedom, democracy, equality, the rule of law and respect for human rights, including the rights of persons belonging to minorities. These values are common to the Member States in a society in which pluralism, non-discrimination, tolerance, justice, solidarity and equality between women and men prevail."*.

[498] United Nations. (n.d.). United Nations and the Rule of Law. Retrieved from What is the Rule. of Law, Retrieved August 2023, from <https://www.un.org/ruleoflaw/what-is-the-rule-of-law/#:~:text=It%20requires%20measures%20to%20ensure,and%20procedural%20and%20legal%20transparency>.

[499] See also: Article 2 of the Treaty on European Union.

The mechanisms of the Rule of Law foster another NATO principle of responsible use: *governability* of AI, since the Rule of Law mechanisms aim to constitute good governance. Good governance is about accountability, transparency, (addressing) liability and contestability. The aim of the mechanisms of the Rule of Law is to produce government that is legitimate and effective. Good governance is about legitimate, accountable and effective ways of obtaining and using public power and resources in the pursuit of legitimate goals.

## 12.3 Separation of Powers: Positive Law & the Rule of Law

Legislative
Executive
Judicial

Figure 3: Separation of powers: independent legislative; executive and judicial power.

Typically, rules and regulations made by the legislative power that concern compliance can be directly applied to real life cases. The use of AI by the executive power in military operations precludes such a straightforward application due to, for example, the current lack of harmonised rules on what is responsible to deploy in the context of the many different specific AI applications (and to various systems). This makes the goal of deploying AI responsibly in military operations even more complicated. Therefore, new guidelines need to be formulated.

This pressing need for guidance, without hindering innovations, in how to cope in a responsible way with upcoming complexities such as the technological change forced by AI applications in military theatres is manifested in many initiatives. Intergovernmental and defence organizations develop their own frameworks for responsible use of AI in the context of military operations. These include and are not limited to NATO,[500] OECD,[501] the European Commission,[502] the US Department of Defense,[503] and the UK Ministry of Defence.[504] Although these frameworks are not 'positive law', these can be informative in how to shape governmental tools if these are brought in relation to the tenets of the Rule of Law and its mechanisms.

Moreover, it is important to take into account that the Rule of Law tenets are also not categorised as 'positive law'. Positive law contains rules and regulations that concerns compliance. However, the Rule of Law and connected mechanisms differ from principles of law that can be applied directly to real life use cases and thus scenarios. The Rule of Law is shaped by many sources, such as: case law; legal doctrine; legal interpretation methods; positive law; rules and regulations; draft rules and regulations and legal theory. Existing Rule of Law mechanisms can be found in, for example, processes of new legislation, or redefining the policies and processes within the boundaries set by existing laws.

In order to foster good governance of AI applications it is necessary to implement additional devices and other means (tools) in existing Rule of Law mechanisms in order to enhance existing processes with additional requirements in order to protect the principles of the Rule of Law. Existing mechanisms can be found in, for example, processes of new legislation, or by redefining the policies within existing laws. All these Rule of Law mechanisms can be enriched with new requirements. These requirements should not only be informed by theories which can be found in the disciplines of law and philosophy (the Rule of Law); insights from the field of (AI) technology should also be included. The *newly found* requirements

[500] Zoe Stanley Lockman, E. H. (2021, October 25). An Artificial Intelligence Strategy for. NATO. Retrieved August 2023, from <https://www.nato.int/docu/review/articles/2021/10/25/an-artificial-intelligence-strategy-for-nato/index.html.>

[501] OECD.AI Policy Observatory. (2023). OECD AI Principles overview. Retrieved August. 2023, from <https://oecd.ai/en/ai-principles>.

[502] European Commission. (2021). Ethics Guidelines for Trustworthy AI. Retrieved August 2023, from <https://ec.europa.eu/futurium/en/ai-alliance-consultation.1.html>.

[503] Board, D. I. (2019). AI principles: recommendations on the ethical use of artificial. intelligence by the department of defense: supporting document. United States Department of Defense.

[504] Ministry of Defence United Kingdom. (2022). Policy paper Ambitious, safe, responsible: our. approach to the delivery of AI enabled capability in Defence. Retrieved August 2023, from <https://www.gov.uk/government/publications/ambitious-safe-responsible-our-approach-to-the-delivery-of-ai-enabled-capability-in-defence/ambitious-safe-responsible-our-approach-to-the-delivery-of-ai-enabled-capability-in-defence#ambitious-delivery-of-capabi>.

can be implemented in existing mechanisms which protect the principles of the Rule of Law.

There are already important basic legal aspects which apply to AI applications in military operations. The International Court of Justice has addressed the issue of the International Humanitarian Law (IHL) principles in the context of the use of weapons and confirmed that these principles of IHL apply *"(…) to all forms of warfare and to all kinds of weapons, those of the past, those of the present and those of the future."*[505] There are currently no harmonised legal requirements for the (pre-)registration of the research design for AI applications, while this could make the development of AI applications in military theatres more transparent. Moreover, the current European draft regulation on AI does not apply since the AI Act Draft excludes military purposes:

> *"(12) This Regulation should also apply to Union institutions, offices, bodies and agencies when acting as a provider or deployer of an AI system.* ***AI systems exclusively developed or used for military purposes should be excluded from the scope of this Regulation where that use falls under the exclusive remit of the Common Foreign and Security Policy regulated under Title V of the Treaty on the European Union (TEU).*** *This Regulation should be without prejudice to the provisions regarding the liability of intermediary service providers set out in Directive 2000/31/EC of the European Parliament and of the Council [as amended by the Digital Services Act]."*[506]

It is not strange that AI for military purposes is excluded from the AI Act Draft, since transparency and secrecy are at odds with each other. Secrets of military AI capacities and purposes cannot be made transparent in the same way as is necessary for use within law enforcement and commercial purposes. Nevertheless, the check by government bodies on the design of military AI is important.

[505] International Court of Justice (1996). Legality Of The Threat Or Use Of Nuclear Weapons. Advisory Opinion Of 8 Jul 1996, para. 78.

[506] AI Act Draft Proposal for a regulation of the European Parliament and of the Council on harmonised rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain Union Legislative Acts (COM(2021)0206 – C9 0146/2021 – 2021/0106(COD)).

This could be done via a (pre-)registration of the research design which ideally also includes details about the choices that were made during the design process of AI applications aimed for deployment in military theatres. Hofman states that requirements on the design of studies need to play a role since:

> *"To qualify research as confirmatory, however, researchers should be required to preregister their research designs, including data pre-processing choices, model specifications, evaluation metrics, and out-of-sample predictions, in a public forum such as the Open Science Framework (https://osf.io). Although strict adherence to these guidelines may not always be possible, following them would dramatically improve the reliability and robustness of results, as well as facilitating comparisons across studies."*[507]

Via the (pre-)registration a set of *(new)* requirements can be implemented in the research design. These requirements are not only informed by theories which can be found in the disciplines of law (compliance), and philosophy (the Rule of Law); also insights from the field of (AI) technology should be included. The *(newly found)* requirements can together be formalised in a pre-registration of the research design. It could become a mandatory step in the acquisition processes. This pre-registration can be linked to an existing legal instrument: the obliged review for new means and/or methods of warfare, as described in Article 36 Additional Protocol I to the Geneva Conventions:

> *"In the study, development, acquisition or adoption of a new weapon, means or method of warfare, a High Contracting Party is under an obligation to determine whether its employment would, in some or all circumstances, be prohibited by this Protocol or by any other rule of international law applicable to the High Contracting Party."*

[507] Hofman, J. M., Sharma, A., & Watts, D. J. (2017). Prediction and explanation in social. systems. Science, 355(6324), 486-488.

Another instrument can be found in the deployment phase in the possible necessity of defining the Rules of Engagement. These instruments could be reshaped in order to deal with the new challenges of the aim to deploy AI applications in a responsible way.

What can go wrong when AI is deployed without a (pre-)registration of the research design? By using a scenario NATO's principles of responsible use can be operationalised in a safe environment. A scenario can demonstrate the necessity that the legal effect, risks and consequences of deploying AI applications are taken into account, prior to (real) deployment, in the acquisition and research design. The C-UAS with AI scenario can give a clear vision on the need to protect the mechanisms of the Rule of Law via mitigating requirements such as a (pre-)registration.

## 12.4 Responsible AI Applications in Military Theatres

In military theatres AI applications can accelerate human decision-making by rapidly translating an overwhelming amount of data into useful information. AI is starting to play a key role in the military domain already. The Russian-Ukraine conflict, for instance, has been called a Living Lab for AI warfare.[508] Ukraine uses AI for target and object recognition using satellite imagery as well as analysis of open source data, like social media content, in order to identify Russian modus operandi.[509] Another example of military usage of AI is swarm intelligence: artificial intelligence, often used in UAS, acting in a coordinated manner without a central control unit.[510]

The opportunities of speeding up decision-making in the military domain can contribute to a military advantage, but danger resides in the processes before deployment, for example via errors in the research design of AI applications which can lead to untrustworthy accuracy rates. These untrustworthy rates can lead to collateral damage, or mistakes which in turn can lead to military disadvantages.

[508] National Defense Magazine, Ukraine A Living Lab for AI Warfare, Retrieved September 2023, from <https://www.nationaldefensemagazine.org/articles/2023/3/24/ukraine-a-living-lab-for-ai-warfare>.

[509] National Defense Magazine, Ukraine A Living Lab for AI Warfare, Retrieved September 2023, from <https://www.nationaldefensemagazine.org/articles/2023/3/24/ukraine-a-living-lab-for-ai-warfare>.

[510] Sentient Digital. MILITARY DRONE SWARM INTELLIGENCE EXPLAINED, Retrieved from September 2023, from <https://sdi.ai/blog/military-drone-swarm-intelligence-explained/>.

The answer to the question whether applying an AI application is disruptive in a good sense, or risky in the bad sense, depends on when, how and where the AI is designed, implemented and deployed.

The ambition of various initiated frameworks[511] is to provide guidelines for responsible use of AI in (military) operational settings. However, a translation from principle to practice is yet to be given. One of the objectives of the interdisciplinary approach in this article is to take the first step in operationalising one set of these principles via the means of a scenario. The goal of the approach is the operationalisation of NATO's principles of responsible use (PRUs). A scenario provides the possibility to reflect on NATO's six principles of responsible use, and to showcase the relation of these principles to the tenets of the Rule of Law and its mechanisms.

Scenarios can be used for a myriad of purposes, ranging from the highly conceptual, strategic level, down to the granular tactical level. A scenario on a conceptual level for strategic explorations can focus, for example, on how and where armed forces should operate in military theatres when there are new AI means and methods of warfare. In this paper a scenario is used to illustrate how an interdisciplinary approach that combines law, philosophy and (AI) technology can be helpful in order to determine what responsible AI entails. The showcased scenario in this paper focuses on how to counter Unmanned Aircraft Systems in a responsible way using AI, and how this effort relates to the six principles of responsible use and the Rule of Law.

## 12.5 NATO's Six Principles of Responsible Use

An international effort for promoting how to shape, amongst other international efforts to formulate principles, AI transparency and/or accountability is the ambition of the NATO Data and Artificial Intelligence Review Board[512] (DARB) which has set the goal of developing a toolkit for Responsible AI Certification Standards building on experiences gained from operational use cases. This

[511] Frameworks of: NATO, OECD, European Commission, the US Department of Defense, and the UK Ministry of Defence.

[512] NATO's Data and Artificial Intelligence Review Board, Retrieved August 2023, from. <https://www.nato.int/cps/en/natohq/official_texts_208374.htm>.

paper supports the ambition of the NATO DARB to develop a toolkit for responsible AI certification standards *(Figure 4)* by examining responsible AI applications in military operations with a focus on the Rule of Law. The Rule of Law tenets and mechanisms foster good governance. The tenets and mechanisms of the Rule of Law are used to reflect upon NATO's six principles of responsible use, and by operationalizing the NATO principles of responsible use (PRUs)[513] via a scenario the interdisciplinary approach is showcased. This approach leads to *(new)* requirements which can contribute to an international certification standard. By doing so, this paper can contribute to the ambitions set by NATO's DARB: "*to govern responsible development and use of AI by helping operationalize PRUs".*[514]

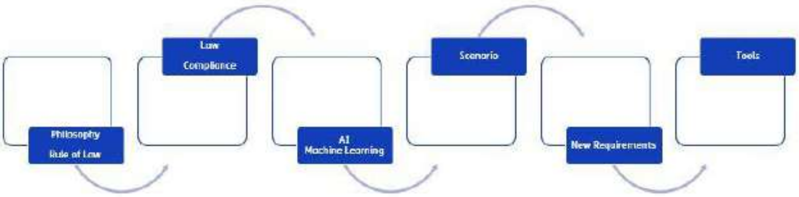


Figure 4: How to use scenarios to find new requirements which can become tools?

Operationalisation can specifically illustrate the upsides of deploying AI, and can also illustrate the downside, namely: how AI applications unintentionally pose a risk to the fundamental norms, values and mechanisms of democratic societies such as the Rule of Law. To these ends, this paper first illustrated how AI applications can be deployed in military theatres with the aim of military advantage, and second how *(new)* requirements can be found. The scenario is fuelled by a Counter Unmanned Aircraft System (C-UAS) use case, where the C-UAS is enhanced with AI applications. By means of this scenario, the paper explains how *(newly found)* requirements can contribute to reshaping existing acquisition and/or research and development processes-both part of existing (international) laws and policies.

[513] The NATO principles of responsible use include: lawfulness; responsibility and accountability; explainability and traceability; reliability; governability; and bias mitigation.

[514] NATO's Data and Artificial Intelligence Review Board, Retrieved August 2023, from. <https://www.nato.int/cps/en/natohq/official_texts_208374.htm>.

## 12.6 Operationalising NATO's Principles of Responsible Use: a scenario of C-UAS with an AI Application

Unmanned Aircraft Systems (UAS), or drones, have influenced modern warfare over the past three decades with applications ranging from real-time intelligence to precision strikes. AI applications may have potential added value to Counter Unmanned Aircraft Systems (C-UAS) for identifying, tracking and defending against these threats but could, without *(new)* requirements in existing acquisition and research design processes, unintentionally violate the tenets of the Rule of Law. To demonstrate how C-UAS with AI applications are related to NATO's principles of responsible use and how these relate to a violation of the Rule of Law, an example of a possible scenario is sketched in this section.

An electro-optic camera system can be enhanced with an AI application in order to automate the distinction of small Unmanned Aircraft Systems (sUAS)[515] from birds. sUAS and birds have similar characteristics and are therefore easily confused with one another during surveillance tasks. The drone-vs-bird challenge is a yearly event where contestants train a deep learning model, which is a specific type of AI, on a given training dataset with the goal of detecting a sUAS in video footage in which birds are also present. The model should trigger when a UAS is detected and give a position estimate of the sUAS while not giving an alert if birds are detected. The AI model in question is a Convolutional Neural Network (CNN), which is a supervised deep learning algorithm, that takes as input images of different types of birds and UAS and yields a classification, bird or UAS, as prediction output.

This AI application is designed to improve identification of potential targets for electro-optic camera systems in a C-UAS; this is only one example of an AI application that could improve the effectiveness of military systems. Yet this same system, if the NATO principles of responsible use have not been fully considered in the design process, could prove a risk to the Rule of Law mechanisms. For example, if the AI application is trained only on video footage of birds and one specific sUAS type, the output of the application is biased

[515] sUAS refers to small UAS, which is a subcategory of NATO Class I UAS with a weight below 15 kgs.

towards one sUAS type and may not be able to recognise other sUAS types. This conflicts with the NATO principles of responsible use *bias mitigation "Proactive steps will be taken to minimize any unintended bias in the development and use of AI applications and in data sets."*[516] Since it is biased towards one sUAS type, and reliability *"AI applications will have explicit, well-defined use cases. The safety, security, and robustness of such capabilities will be subject to testing and assurance within those use cases across their entire life cycle, including through established NATO and/or national certification procedures."*[517] As birds cannot be reliably distinguished from other sUAS types.

Furthermore, erroneous sUAS identifications by the system could lead to potentially erroneous actions or decisions during operations and military disadvantages. If the NATO principles of responsible use *responsibility* and *accountability (Figure: 2) ("AI applications will be developed and used with appropriate levels of judgment and care; clear human responsibility shall apply in order to ensure accountability."*[518]*)* are taken into account insufficiently during the research design, it becomes unclear who or what is to blame for errors or how these could be corrected. These errors, since the biased actions can lead to mistakes in the kill chain, can accidentally give away the location of the C-UAS with AI application. This can lead to attacks, conducted by the enemy against the military material, leading to a military disadvantage.

Consequently, the (AI) technology may violate the mechanisms of the Rule of Law i.e. *"procedural and legal transparency."*[519] Since accountability questions cannot be addressed.

[516] NATO's Data and Artificial Intelligence Review Board, Retrieved August 2023, from. <https://www.nato.int/cps/en/natohq/official_texts_208374.htm>.

[517] NATO's Data and Artificial Intelligence Review Board, Retrieved August 2023, from. <https://www.nato.int/cps/en/natohq/official_texts_208374.htm>.

[518] NATO's Data and Artificial Intelligence Review Board, Retrieved August 2023, from. <https://www.nato.int/cps/en/natohq/official_texts_208374.htm>.

[519] United Nations. (n.d.). United Nations and the Rule of Law. What is the Rule. of Law, Retrieved August 2023, from <https://www.un.org/ruleoflaw/what-is-the-rule-of-law/#:~:text=It%20requires%20measures%20to%20ensure,and%20procedural%20and%20legal%20transparency>.

## 12.7 From Risks and Advantages to (new) Requirements

The above example illustrates issues that may arise for C-UAS enhanced with AI applications when the NATO principles of responsible use *bias mitigation*, reliability and responsibility and accountability are not considered. To adhere to the *bias mitigation* principles of responsible use for this example, a first requirement is that the dataset consists of a wide range of bird species and sUAS types in a relevant ratio. Here, that means a large collection of different birds and sUAS video footage that reflects different species and types.

The amount of footage per specie or type should be in balance with all other types: a dataset containing just one video of sparrows and a hundred videos of starlings is imbalanced. Secondly, an audit of the dataset by (external) experts would be required to verify that it adheres to the first requirement. To adhere to the principle of *reliability (Figure: 2)*, the C-UAS with AI should be tested in many different settings. To that end, the (AI) technology should be tested on a wide range of birds and sUAS in different landscapes (urban, desert, sea etc.) during the training phase. For example, if one video contains six different sUAS types and one bird the C-UAS should still distinguish the bird from the sUAS.

It is very important to verify and validate, based on the test- and validation data, the performance of the AI model. This needs to be done in a way that it becomes safe for usage. In addition, since a CNN is a deep learning model, the output of the application can not be easily traced back to the input. For example: which features contributed the most to this prediction? To adhere to the principle of *explainability* and *traceability*, the C-UAS with AI should be capable of explaining why it came to a specific prediction. If the AI in question is opaque by nature, such as a CNN, a post-hoc explanation should be given that clarifies, at least partially, which feature contributed the most to the prediction.

As generation of the dataset is done early in the research and development, and testing in the validation and verification phase, the requirements can be applied throughout the entire acquisition and/or development process. To verify the requirements for adherence to the NATO principles of responsible use in relation to the process, testing can be done using operational tests, serious wargaming or digital twins.

Overall, this example shows that active mitigation of risks and errors starts early in the acquisition and/or development process, and hence where the protection of the Rule of Law begins. By integrating the requirements, as translation of the NATO principles of responsible use into the Rule of Law mechanisms, the fundamental norms and values of democratic societies can be protected and the Rule of Law strengthened.

This is just one example of a component of an operational scenario that could be used for clarifying what is at stake when applying AI-enriched technologies for end-users, policymakers, regulatory authorities, researchers, and industry. Any specification of a scenario would depend on the examined AI application and its respective system, as well as the main challenge for the involved stakeholders. Clarifications resulting from the full scenario can provide insight to the legislative, executive and judicial powers on the problems that may occur during deployment when AI is not designed and/or procured in a responsible way, and how this may affect the fundamental norms and values of the democratic society.

## 12.8 Conclusion and Future Study

This paper illustrated how interdisciplinary research, in the form of a scenario, can be a useful approach in taking the first steps to operationalise the NATO principles of responsible use. Bringing together the lenses of the disciplines law, philosophy (the Rule of Law) and (AI) technology in operational scenarios clarifies the risks and military advantages. This clarification can lead to the identification of *(new)* requirements, which can be seen as tools to strengthen the mechanisms of the Rule of Law and hereby democratic societies.

Stakeholders need tools in order to enable the responsible use of AI. When the impact on the norms and values of democratic societies is not clarified to stakeholders of the value chain of partners via a scenario in an applied setting and *(new)* requirements are not used on a tactical and conceptual level as tools to be implemented by end-users, policymakers, regulatory authorities, researchers, and industry in the Rule of Law mechanisms, it might be the case that the AI application cannot be deployed in a responsible nor lawful way.

Verification and validation of these *new* requirements is required. The verification and validation can be developed via tests in (close to) real life environments, for example in operational tests, serious gaming, and digital twins.

Finally, the *(newly found)* requirements are the tools which can be presented in a demystified way to decision makers and other relevant stakeholders in how these can inform the reshaping of legislation, certification, and policy-guidelines in existing processes.

## 12.9 Appendix A

List of Key Definitions

| | |
|---|---|
| **AI** | Artificial Intelligence |
| **AI Act Draft** | European legislation on harmonised rules on Artificial Intelligence |
| **AI application** | Application of AI (this can be Machine Learning) in systems |
| **CNN** | Convolutional neural network |
| **C-UAS** | Counter Unmanned Aircraft Systems |
| **DARB** | NATO's Data AI Review Board |
| **EDT** | Emerging Disruptive Technology |
| **NATO** | North Atlantic Treaty Organization |
| **PRUs** | NATO's Principles of Responsible Use of AI |
| **Requirements** | Technical requirements which can become tools of good governance |
| **Rule of Law** | Is shaped by various sources, such as: case law; legal doctrine; legal interpretation methods; positive law; rules and regulations; draft rules and regulations and legal theory. |
| **sUAS** | Small Unmanned Aircraft Systems |
| **(AI) Technology** | AI technologies and/or systems with applied AI applications |
| **Rule of Law tenets** | *accountability, transparency, contestability* |
| **mechanisms** | *processes of rules and regulations; case law; policies; etc.* |

# PART 4

## Asian and African Perspectives on AI Governance

# 13 Operationalizable Accountability of (Generative) AI: Towards the Chinese AI Law?

**Wayne Wei Wang, PhD Candidate, Faculty of Law, University of Hong Kong & Fellow-in-Rio at Fundação Getulio Vargas Rio Law School (FGV Direito Rio).**

**Yue Zhu, Assistant Professor, School of Law, Tongji University & Assistant Research Fellow, Shanghai Collaborative Innovation Center of Artificial Intelligence for Social Governance.**

## Abstract

The policy paper elucidates the disparate conceptualizations of Artificial Intelligence (AI) accountability among various stakeholders and actors, thereby facilitating an informed discussion about the ambiguity and implementability of normative frameworks governing AI. Specifically, the paper examines ongoing global regulatory initiatives pertinent to Generative AI, situated within the realm of data protection legislation and reveals how these piecemeal strategies have been integrated into a composite methodology pending the enactment of formal statutes. Employing China as a jurisdictional case study, the policy paper scrutinizes the evolution of regulatory landscapes with respect to ethics, content security, and data protection. A comparative analysis is executed between the in-force Interim Measures for Generative AI and the non-binding Chinese Model AI Law, newly proposed by Chinese academia in 2023. The paper culminates in positing that the operationalization of AI accountability can only be achieved through an institutional architecture that delineates specific obligations and liabilities, including but not limited to retention, disclosure, and mutual assistance. Furthermore, it advocates for a calibrated balance between adaptability and regulatory predictability to facilitate agile, realistic, and implementable accountability.

## 13.1 Introduction

The feasibility of artificial intelligence's accountability was once deemed elusive, intrinsically intertwined with its multifaceted sub-concepts, such as explainability, accuracy, reliability, and robustness, to name a few. A salient illustration is the European Union's AI Act, which once, through Article 13 of its 2021 proposal,[520] engendered innovation-vs.-regulation discussions regarding the intricate delineations between explainability and interpretability (Grady, 2022).[521] While some models might elude full interpretation, an array of model-agnostic techniques exist to elucidate their operations as part of the traditional machine-learning explanation (Du et al., 2019).[522] Moreover, in specific pivotal AI applications, such as within adversarial contexts, mechanisms like post-hoc-explanation-enabled explainability don't always function optimally (Bordt et al., 2022).[523]

In addition to the technical and endogenous challenges above, the inherent intricacy and autonomy of AI systems pose dilemmas in attributing responsibility for their decisions, which occasionally fall prey to gaming – manifesting through the employment of proxies and estimators within decisional processes (Bambauer & Zarsky, 2018).[524] In contrast to human agents, AI entities lack moral agency and sentience, complicating the attribution of culpability or their amenability to account (Coeckelbergh, 2020).[525] This demands reasonable clarity regarding the liable parties when AI systems err or inflict damage. An additional impediment to realizing AI accountability is the swift trajectory of technological progress. The evolution of AI technologies is meteoric, with (proposed) regulatory and ethical paradigms grappling to remain contemporaneous in a race to regulate or govern it (Bradford, 2023).[526] The operationalization of these paradigms and ensuring universal compliance across diverse

[520] Article 13 of the 2021 Proposal of the EU AI Act: "High-risk AI systems shall be designed and developed in such a way to ensure that their operation is sufficiently transparent to enable users to interpret the system's output and use it appropriately.".

[521] Grady, P. (2022, August 31). The EU Should Clarify the Distinction Between Explainability and Interpretability in the AI Act. Center for Data Innovation. <https://datainnovation.org/2022/08/the-eu-should-clarify-the-distinction-between-explainability-and-interpretability-in-the-ai-act/>.

[522] Du, M., Liu, N., & Hu, X. (2019). Techniques for interpretable machine learning. Communications of the ACM, 63(1), 68–77. <https://doi.org/10.1145/3359786>.

[523] Bordt, S., Finck, M., Raidl, E., & von Luxburg, U. (2022). Post-Hoc Explanations Fail to Achieve their Purpose in Adversarial Contexts. *2022 ACM Conference on Fairness, Accountability, and Transparency*, 891–905.<https://doi.org/10.1145/3531146.3533153>.

[524] Bambauer, J., & Zarsky, T. (2018). The Algorithm Game. Notre Dame Law Review, 94(1), 1–48.

[525] Coeckelbergh, M. (2020). Artificial Intelligence, Responsibility Attribution, and a Relational Justification of Explainability. Science and Engineering Ethics, 26(4), 2051–2068. <https://doi.org/10.1007/s11948-019-00146-8>.

[526] Bradford, A. (2023, June 27). The Race to Regulate Artificial Intelligence. Foreign Affairs. <https://www.foreignaffairs.com/united-states/race-regulate-artificial-intelligence>.

sectors and jurisdictions constitutes a formidable challenge, with diverging stakeholder-specific and actor-oriented ethical perceptions (Jobin et al., 2019).[527]

Generative AI, especially Large Language Models (LLMs) like ChatGPT, has already attracted global regulatory attention, with data protection authorities acting swiftly. After ChatGPT was launched on November 30, 2022, it initially faced a ban in Italy in March and became available again in April after compliance campaigns (Lomas, 2023).[528] The Italian Data Protection Authority, known as the Garante, argued that ChatGPT could potentially be in violation of the European Union's General Data Protection Regulation (GDPR). The Garante chose to impose an immediate, temporary limitation on the processing of data related to ChatGPT, pending further investigation, in accordance with Article 58(2)(f) of the GDPR.[529] Similar investigations were followed by France's CNIL, Spain's data protection agency and Japan's privacy watchdog (Parodi & Orusov, 2023).[530] In the meantime, some have argued that the use of personal data for training purposes has legality flaws that will persist until generative AI technologies, like large language models, become fully compatible with data protection (Belli, 2023).[531]

This paradigm of sector-specific (e.g., data protection) accountability also constitutes the developmental trajectory of artificial intelligence regulation in China. Commencing with ethics-oriented soft constraints, transitioning to sectoral hybrid governance, and culminating in specialized oversight of generative AI, the understanding of accountability among various stakeholders has been dynamically evolving. This evolution is epitomized by the Expert Draft Proposal formulated in 2023, encapsulated in the Artificial Intelligence Law (Model Law) in China.

[527] Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. Nature Machine Intelligence, 1(9), Article 9. <https://doi.org/10.1038/s42256-019-0088-2>.

[528] Lomas, N. (2023, April 28). ChatGPT resumes service in Italy after adding privacy disclosures and controls. TechCrunch. <https://techcrunch.com/2023/04/28/chatgpt-resumes-in-italy/.

[529] The Garante's decision can be found at <https://www.garanteprivacy.it/web/guest/home/docweb/-/docweb-display/docweb/9870847.

[530] Parodi, A., & Orusov, A. (2023, August 22). Governments race to regulate AI tools. Reuters. <https://www.reuters.com/technology/governments-race-regulate-ai-tools-2023-08-22/.

[531] Belli, L. (2023, May 25). Why ChatGPT does not comply with the Brazilian Data Protection Law. MediaNama. <https://www.medianama.com/2023/05/223-chatgpt-brazilian-data-protection-law-ai-regulation/>.

## 13.2 The Hybrid Regulatory Approach

In the contemporary milieu where global AI governance frameworks remain in a nascent stage of discursive proposal, notable stakeholders such as the OpenAI's CEO have articulated a welcoming stance towards regulatory scrutiny of algorithmic safety during congressional hearings in the United States (Kang, 2023).[532] Such pronouncements posit the notion of a unified global framework for AI governance, which somehow may only reach consensus after quite a long time, allowing enterprises to execute regulatory arbitrage (Pollman, 2019).[533] Simultaneously, as the objective remains to maintain technological neutrality yet mandates the deployment of technologies for effective implementation, the techno-neutral principles inherent in extant data protection legislation, for instance, in the EU's GDPR (Wong & Henderson, 2019),[534] are encountering challenges.

These laws presently serve as a provisional regulatory mechanism for artificial intelligence. This scenario concurrently underscores the lacuna in effective enforcement tools and instruments that regulators possess with respect to issues like disinformation/misinformation, defamation, and intellectual property infringement by generative AI. Contrastingly, there are more than 160 jurisdictions globally that have already enacted data protection legislation.[535] With regard to generative AI and its implications for the collection, utilization, and sharing of personal data—ChatGPT serving as an illustrative case in point—significant legal deficiencies are evident (Burgess, 2023).[536]

[532] 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>.

[533] Pollman, E. (2019). Tech, Regulatory Arbitrage, and Limits. *European Business Organization Law Review*, *20*(3), 567–590. <https://doi.org/10.1007/s40804-019-00155-x>.

[534] Wong, J., & Henderson, T. (2019). The right to data portability in practice: Exploring the implications of the technologically neutral GDPR. International Data Privacy Law, 9(3), 173–191. <https://doi.org/10.1093/idpl/ipz008>.

[535] As per the statistics by Privacy Law & Business, by Feb 2023, there had been 162 national laws and 20 Bills that were relevant to privacy or data protection. See <https://www.privacylaws.com/reports-gateway/articles/int181/int181_2023/>.

[536] Burgess, M. (2023). ChatGPT Has a Big Privacy Problem. *Wired*. <https://www.wired.com/story/italy-ban-chatgpt-privacy-gdpr/>.

## 13.3 “Soft” Ethics

Pertaining specifically to the governance landscape in China, the initial approach largely mirrored that of the majority of nations and conformed to international trends, wherein ethics, with parallel but progressive operation, served as a form of soft law. To elucidate, the Ministry of Science and Technology inaugurated the Next Generation Artificial Intelligence Governance Committee in 2019 (Zeng, 2020).[537] Subsequently, on June 17 of the same year, the Committee promulgated the “Principles for the Governance of New Generation Artificial Intelligence – Developing Responsible Artificial Intelligence.” (Next Generation Artificial Intelligence Governance Committee, 2019).[538]

The “Guidelines on Strengthening the Governance of Technological Ethics”, later proposed in 2021 and launched in 2022, define precepts such as the enhancement of human well-being, reverence for the sanctity of life, adherence to fairness and justice, prudent risk management, along with the sustenance of openness and transparency, as cardinal doctrines of technological ethics, whilst allocating responsibilities to innovative stakeholders, inclusive of researchers, in the execution of technological ethics governance (General Office of the CCP Central Committee & State Council Office, 2022).[539]

Specifically, the 2021 “Ethical Standards for the New Generation of Artificial Intelligence”, cogently stipulates rudimentary ethical norms for artificial intelligence, whilst concurrently suggesting an assortment of management standards, research and development criteria, supply norms, and usage guidelines for AI implementations, with the overarching objective of infusing ethical considerations

[537] Zeng, Y. (2020). 推┌新一代人工智能┌理┌治理 [Promoting the ethics and governance of the new generation of artificial intelligence]. ST Daily. http://www.stdaily.com/index/kejixinwen/2020-06/04/content_952415.shtml.

[538] Next Generation Artificial Intelligence Governance Committee. (2019). 新一代人工智能治理原┌——┌展┌┌任的人工智能 [Principles for Next Generation AI Governance—Developing Responsible AI]. <https://www.most.gov.cn/kjbgz/201906/t20190617_147107.html>.

[539] General Office of the CCP Central Committee & State Council Office. (2022). ┌于加强科技┌理治理的意┌ [Guidelines on Strengthening the Governance of Technological Ethics]. <https://www.gov.cn/zhengce/202212/content_6688372.htm>.

into the comprehensive lifecycle of AI (Next Generation Artificial Intelligence Governance Committee, 2021).[540]

## 13.4 Content Security

China’s early policy documents mapped out the embryonic form of governing algorithmic recommendations and deep synthesis. As early as 2020, in its “Outline for Establishing a Rule-of-Law-Based Society (2020–2025)”, as a declarative party document with no legal effects, put forth by the CCP Central Committee, China called for initiatives to counteract recommendation algorithms and deepfakes (Central Committee of the Chinese Communist Party, 2020).[541] Subsequently, the Cyberspace Administration of China, in conjunction with other entities, collectively promulgated the “Guiding Opinions on Strengthening Comprehensive Governance of Internet Information Service Algorithms” in the year 2021 (Cyberspace Administration of China et al., 2021).[542]

By also integrating the 2021 Provisions on the Management of Algorithm Recommendation for Internet Information Services (CAC et al., 2021),[543] an administrative regulation, China established a comprehensive governance system for algorithms gradually; with regard to five types of algorithms, namely 1) generation and synthesis, 2) personalized push, 3) sorting and selection, 4) retrieval and filtering, and 5) scheduling and decision-making, those requirements for the establishment of an internal management system, evaluation and verification, public notice, user self-management, algorithm registry, etc., are set forth.

[540] Next Generation Artificial Intelligence Governance Committee. (2021). 新一代人工智能┌理┌范 [Ethical Standards for the New Generation of Artificial Intelligence]. <https://www.most.gov.cn/kjbgz/202109/t20210926_177063.html>.

[541] Central Committee of the Chinese Communist Party. (2020). 法治社┌建┌┌施┌要 (2020-2025 年) [Outline for Establishing a Rule-of-Law-Based Society (2020–2025)]. <https://www.gov.cn/zhengce/2020-12/07/content_5567791.htm.

[542] Cyberspace Administration of China, Central Propaganda Department, Ministry of Education, Ministry of Science and Technology, & Ministry of Industry and Information Technology. (2021). ┌于加强互┌┌信息服┌算法┌合治理的指┌意┌ [Guiding Opinions on Strengthening Comprehensive Governance of Internet Information Service Algorithms]. <http://www.cac.gov.cn/2021-09/29/c_1634507915623047.htm>.

[543] AC, MIIT, MPS, & SAMR. (2021). 互┌┌信息服┌算法推荐管理┌定 [Provisions on the Management of Algorithm Recommendation for Internet Information Services]. <https://www.gov.cn/zhengce/zhengceku/2022-01/04/content_5666429.htm.

China has implemented regulations on deep synthesis technology, which refers to the utilization of generative synthetic algorithms such as deep learning and virtual reality to produce web-based information, including text, images, audio, video, and virtual scenes, as outlined in the 2022 Provisions on the Management of Deep Synthesis Internet Information Services (CAC et al., 2022).[544]

## 13.5 Data Protection

As briefed above, on one hand, Article 24 of the Personal Information Protection Law (PIPL),[545] can be understood as adhering closely to FAccT principles, emphasizing transparency in automated decision-making, ensuring fairness in the outcomes, advocating for user autonomy and consent, requiring accountability in operations, and granting the right to refuse automated decisions that significantly impact individual rights and interests.

Implying that the algorithms and data sets behind the decision-making must be comprehensible to both the subjects and auditors, thereby facilitating scrutiny and understandability of the process, the article mandates the need for equitability in the system, prohibiting "unreasonable differential treatment" of individuals in trading conditions, akin to preventing discriminatory or unjust practices. At the same time, those using automated methods for information dissemination or commercial sales to individuals are required by the article to provide opt-out mechanisms, ensuring individual autonomy, and fostering informed consent. Individuals should have the right to demand an explanation and should have the option to refuse to be solely subject to automated decision-making processes if such decisions have a substantial influence on their rights and interests. This resonates with human oversight and the option for human intervention in consequential decision-making scenarios.

That said, pursuant to Article 24 of the Personal Information Protection Law (PIPL), the statute predominantly encompasses:

[544] CAC, MIIT, & MPS. (2022). 互┌┌信息服┌深度合成管理┌定 [Provisions on the Management of Deep Synthesis Internet Information Services]. <https://www.gov.cn/zhengce/zhengceku/2022-12/12/content_5731431.htm.

[545] See the English Translation of the PIPL for reference at <https://digichina.stanford.edu/work/translation-personal-information-protection-law-of-the-peoples-republic-of-china-effective-nov-1-2021/.

**1.** an individual's entitlement to equitable commerce or fair trade (i.e., "forbidding unreasonable differential treatment of individuals in trading conditions such as trade price") and

**2.** adherence to their autonomous selection prerogative (i.e., "the option not to target an individual's characteristics or a convenient method to refuse"),

**3.** examined through the lens of algorithmic transparency (i.e., "in cases of a major influence on the rights and interests, the right to require personal information handlers to explain the matter, and the right to refuse that personal information handlers make decisions solely through automated decision-making methods").

On the other hand, generally, Article 44 of the PIPL ensures that PI subjects shall be adequately informed of the relevant handling activities and shall have the right to restrict or refuse the handling of their personal information by the enterprise; at the same time, Article 48 states that individuals have the right to ask companies to explain their rules for handling personal information.

Article 23 and Article 29 of the PIPL mandate rigorous consent protocols for personal information handlers, particularly concerning the transfer of personal data to another handler and the management of sensitive information, including that of minors. According to Article 23 of the PIPL, if a PI handler provides personal information processed by it to another one, the handler shall inform the individual of the recipient's name, contact information, the purpose and method of the handling, and the type of personal information, and shall obtain the individual's separate consent. According to Article 29 of the PIPL, the handling of sensitive personal information shall be subject to the individual's separate consent. In most cases, personal information of children under the age of 14 (inclusive) and information relating to the privacy of natural persons is sensitive personal information. When the handler wants to do automated data processing, it should thus consider those criteria of (separate) consent.

In addition, specific regulatory frameworks such as Article 18 of the E-Commerce Law and Article 24 of the PIPL, along with the 2021 Provisions on the Management of Algorithm Recommendation for Internet Information Services (i.e. Articles 2, 10, 17, and 21), collectively stipulate stringent guidelines for user profiling, offering options for non-targeted content, mandating fair trade and algorithmic transparency, and providing mechanisms for an immediate cessation of user profiling and tag deletion.

## 13.6 Accountability in the Chinese AI Law (Model Law v.1.0)

China took a significant step in pursuing AI accountability by initially proposing a draft regulation on Generative AI on April 11, 2023, and later formalizing it on July 13, 2023. This regulation, known as the Interim Measures for the Administration of Generative Artificial Intelligence Services, aims to strike a balance between technological growth and compliance obligations. It emphasizes a nuanced approach to governance, focusing on data compliance, intellectual property, and content security, among other issues.

Simultaneously, the Artificial Intelligence Law – Model Law v.1.0 (Expert Draft Proposal), commonly referred to as the "Model Law," introduces the principle of accountability. It outlines a complex framework that categorizes entities involved in the AI value chain into developers, providers, and users, each with varying degrees of responsibility. The law also stipulates specific obligations for these entities, such as retention, disclosure, and mutual assistance, to ensure a transparent and accountable AI ecosystem. The Model Law goes beyond mere compliance to advocate for proactive, embedded accountability in AI design.

## 13.7 The Interim Measures

On April 11, 2023, China proposed a specific regulation on Generative AI – Measures for the Administration of Generative Artificial Intelligence Services (Draft for Comment) (CAC, 2023),[546] which discussed some core issues such as 1) Data-related Compliance: Infringement of personal information and/or privacy, and trade secrets; 2) Intellectual Property: Breach of others' copyrights as regards training data; 3)

[546] CAC. (2023, April 11). 生成式人工智能服┌管理┌法（征求意┌稿）[Measures for the Administration of Generative Artificial Intelligence Services (Draft for Comment)]. <http://www.news.cn/politics/2023-04/11/c_1129511663.htm>.

Content Security: Dissemination of disinformation or misinformation, manipulation of public opinion, and engagement in cognitive conflict.

Subsequently, on July 13, 2023, the CAC, together with six ministries, formally issued the Interim Measures for the Administration of Generative Artificial Intelligence Services (hereinafter Interim Measures) (CAC et al., 2023),[547] which took effect in August of 2023. The Measures optimize the compliance obligations of generative AI service providers, leaving a certain buffer for companies between pursuing growth and meeting compliance obligations.

In particular, the Measures propose that the State adhere to the principles of equal emphasis on development and safety, promotion of innovation and rule-of-law, take effective measures to encourage the innovation and development of generative AI, and implement inclusive, prudent, and categorized and graded supervision of generative AI services. For instance, it takes the Law on Scientific and Technological Progress as its superseding law, emphasizing its core concept of promoting scientific and technological progress in AI services.[548]

## 13.8 The Model Law

Commissioned by the Chinese Academy of Social Sciences (CASS)'s National Conditions Research Project, a working group on AI Ethical Review and Regulatory System proposed the *Artificial Intelligence Law – Model Law v.1.0 (Expert Draft Proposal)* in 2023, hereinafter referred to as the "Model Law".[549] Specifically, the accountability principle in Article 7 of the Model Law encompasses three distinct aspects of normativity.

First, it delineates and categorizes entities suitable for assuming responsibility within the increasingly intricate AI industry/value chain. Second, by stipulating retention, disclosure, and mutual assistance obligations tailored to diverse entities, it lays the groundwork for

[547] CAC, NDRC, ME, & MST. (2023, July 13). 生成式人工智能服┌管理┌行┌法 [Interim Measures for the Administration of Generative Artificial Intelligence Services]. <http://www.cac.gov.cn/2023-07/13/c_1690898327029107.htm>.

[548] See Art.1, Interim Measures.

[549] See the English Translation of the Model Law for reference at <https://digichina.stanford.edu/work/translation-artificial-intelligence-law-model-law-v-1-0-expert-suggestion-draft-aug-2023/>.

achieving AI accountability through appropriate legal, institutional, and technological frameworks. Lastly, from a more extended perspective, it advocates for various AI entities to proactively research and implement more accountable technological architectures in a robust, anticipatory, and embedded manner while elucidating their commitment to fulfilling societal expectations of responsibility and detailing the manner of such undertakings.

Given the complexity and heterogeneity intrinsic to the AI industry and value chain, the precondition for achieving AI accountability rests on the identification and categorization of specific types of responsible entities. Evidently, it is untenable to impose uniform responsibilities on disparate actors, such as "gatekeepers," leading-edge AI startups, contributors within open-source communities, academic institutions, research organizations, or philanthropic entities vis-à-vis commercial enterprises. Traditional criteria for apportioning liability, including the capability to control technology domains, proficiency in risk identification and mitigation, financial/profiting capacity, and the accrual of commercial benefits, remain highly relevant in AI legislation. Entities with greater control over technology, superior aptitude in risk recognition and remediation, deeper financial resources, and those deriving commercial advantages from AI might accordingly bear heightened responsibilities.

Predicated on the above, Article 71 of the Model Law disaggregates the intricate and heterogeneous AI industry and value chain into three distinct categories of entities—developers, providers, and users.

Developers are those solely engaged in research and development activities such as algorithmic design, model optimization, and testing deployment. Providers are entities offering AI for commercial purposes or serving an indefinite public. The responsibility vested in developers is comparatively lighter than that allocated to providers. Worth noting is that if a singular entity partakes in both development and provision activities, it should be classified under the more heavily accountable category of "provider" to prevent evasion of responsibility from providers to developers, thereby impeding the realization of accountability. Users are the entities deploying AI, possessing a degree of control and remediation capabilities over AI's outputs and associated risks; they, too, may assume responsibility. The hierarchical distribution of technical control between providers and users varies based on the specific context, and consequently, the apportionment of their responsibilities must be contextually contingent.

To be more precise, the operationalization of AI accountability necessitates the stipulation of retention, disclosure, and mutual assistance obligations tailored to varying types of entities. Without authentic, comprehensive, and accurate information regarding how AI is developed, provided, and utilized, it becomes arduous to ascertain causality, adjudicate culpability, and allocate responsibility when risks or damages arise due to AI. Such a vacuum of information culminates in deleterious outcomes both ex-ante and ex-post. In the pre-event phase, if the entities causing danger or committing errors are not fully accountable, a "moral hazard" situation arises, leading to suboptimal precautionary measures. In the post-event phase, the absence of accountability or an inequitable distribution of responsibility—either too lenient or too stringent—ensues in manifest injustices. To appropriately attribute and hold accountable, the "Model Law" mandates that developers, providers, and users retain essential information and disclose or provide it to other relevant entities when requisite.

The Model Law furnishes explicit stipulations across three dimensions—retention, disclosure, and mutual assistance. Article 33, in conjunction with relevant clauses, mandates that AI developers and providers adhere to the legislative requirements for record-keeping and retention of technical documents to "ensure (AI's) traceability." This predominantly embodies the retention of technical documents related to quality control, risk assessment, and security vulnerabilities. Article 35, along with associated provisions, codifies the transparency obligations incumbent upon developers and providers. Additionally, Articles 35 and 42, among others, delineate mutual assistance obligations between various types of entities, particularly between developers and providers as well. For instance, the concluding clause of Article 35 prescribes that developers are obliged to assist providers in publicly disclosing the fundamental principles, intended objectives, and primary operational mechanisms of AI products and services.

Last but not least, analogous to other foundational principles such as privacy, fairness, and environmental sustainability, accountability, in its ideal form, should be realized through an embedded design approach. This implies that accountability ought to be proactive rather than reactive, preventative rather than remedial, and positive-sum rather than zero-sum (Cavoukian, 2011).[550] Only when accountability is intrinsically assured at the design level of AI can we confidently assert that societal pursuits concerning this cornerstone principle of AI governance have been successfully attained. Absent such intrinsic assurances in technological design, residual risks will perpetually linger. The Model Law also apparently aims to encourage AI developers, providers, and users to explicitly delineate whether and how they assume the societal responsibilities expected of them, thereby facilitating the establishment of high-calibre, trust-based governance.

In the examination of accountability mechanisms, whether oriented towards technological design parameters or necessitating an explicit articulation of such accountability, it is apparent that both approaches could entail a specific set of obligations and may introduce significant technical impediments. To navigate these complexities, the Model Law adopts a circumspect approach in promulgating such stipulations. This orientation aligns, to a degree, with the principles outlined in Article 4a of the draft AI Act as proposed by the European Parliament.[551] On one axis, the law incorporates principle-based provisions, thus affording the necessary latitude to adapt to forthcoming technological advancements. On the complementary axis, these principle-based tenets are confined to specialized exceptional circumstances and are typically operationalized at the level of specific rules or norms. This equilibrium between adaptability and regulatory predictability is imperative for the ongoing assurance of effective accountability within the intricate, variegated, and perpetually evolving landscape of the artificial intelligence value chain.

## 13.9 Conclusion

The policy paper commences by scrutinizing the historical construct of AI accountability, a concept that was previously elusive due to

[550] Cavoukian, A. (2011). *Privacy by Design: The 7 Foundational Principles*. <https://iab.org/wp-content/IAB-uploads/2011/03/fred_carter.pdf>.

[551] See the legislative version referred to as the European Parliament's Version, in particular, Amendment 213 at <https://www.europarl.europa.eu/doceo/document/TA-9-2023-0236_EN.html.>

the intricate interconnection of notions encompassing explainability, precision, dependability, and resilience. It delves into the intricate challenges presented by the autonomy and complexity inherent in AI systems, demanding unequivocal precision in the allocation of responsibility for their decision-making. The rapid trajectory of technological progression further compounds the intricacies involved in establishing efficacious regulatory and ethical paradigms, given that AI is undergoing unprecedented evolution.

China's transition towards a hybrid regulatory paradigm serves as an illustrative archetype underscoring the requisite for a cohesive global framework dedicated to the governance of AI, all the while acknowledging the obstacles in maintaining technological neutrality within prevailing statutes. The contemporary approach adopted by China entails the integration of its hybrid regulatory strategy through the prism of content security, data protection, and sector-specific regulation pertaining to Generative AI, allegedly harmonizing expansion imperatives with responsibilities tied to adherence. More explicitly, the Model Law introduces the principle of accountability, stratifying the entities along the AI value chain and enunciating duties relating to responsibilities, retention, disclosure, and mutual assistance, all with the intent of fostering a transparent and responsible AI ecosystem.

Encompassing technical, ethical, and regulatory dimensions of AI accountability, it highlights the evolving responses of jurisdictions such as China to address these challenges, emphasizing accountability principles, tailored obligations, and proactive technological design. As AI continues its metamorphic influence on societies and industries, the pursuit of implementable, or operationalizable accountability remains an exigent imperative, necessitating a nuanced equilibrium between adaptability and regulatory predictability to assure efficacious and agile governance within the dynamic AI landscape.

# 14 Seeking Policy, Technical and Operational Transparency in AI Systems: A Case Study of India's Digi Yatra Project

**Smriti Parsheera, PhD candidate at the Indian Institute of Technology Delhi**

## Abstract

Transparency is widely recognized to be one of the basic principles of good governance of artificial intelligence (AI). This paper discusses the *why* and *how* of transparency obligations, as articulated in the AI governance discussions in India and in select international principles. It argues that the need for transparency permeates through the lifecycle of an AI project and identifies the policy layer, the technical layer and the operational layer as the key sites for fostering the transparency in any AI project. It then studies India's Digi Yatra project, a system for biometric identity verification at airports, to examine how the project fares on transparency across these three identified layers. The paper points to certain gaps in the discharge of transparency obligations in connection with the Digi Yatra project and its lessons for AI transparency, particularly in the context of large public projects.

## 14.1 Introduction

Transparency in the deployment and operation of AI systems has emerged as one of the universally accepted principles of AI governance. Its goal is to enable individuals to understand when and why AI-based decision making is taking place and to be able to hold the relevant actors to account. Much of the AI transparency debate falls under the realm of what Andrada et al., 2022 describe as 'reflective transparency' – seeking insights into the process of AI decision making and opening up its details or constituents to further deliberation.[552] In the process, the principle of transparency

[552] The authors distinguish this from the philosophical concept of 'transparency-in-use' that is achieved when a technology becomes transparent to the user in an experiential sense through skillful interactions with it. See Andrada G, Clowes R.W., Smart P.R. (2023). Varieties of transparency: exploring agency within AI systems. AI & Society. 38. p. 1321–1331. Retrieved from <https://link.springer.com/article/10.1007/s00146-021-01326-6>.

is often linked with two other key principles, of explainability and accountability. This paper discusses the ways in which the link between transparency, explainability and accountability comes through in various international AI principles and in the AI strategy adopted by India. It chooses the AI principles adopted by the United Nations Educational, Scientific and Cultural Organization (UNESCO) and the Organisation for Economic Co-operation and Development (OECD) as relevant examples.

The process of developing India's principles for responsible AI began with a 2018 discussion paper issued by the government's official think tank, the NITI Aayog.[553] In a subsequent publication, the NITI Aayog laid out its recommended Principles for Responsible AI.[554] It identified seven broad principles for the responsible management of AI. These are safety and reliability, equality, inclusivity and non-discrimination, privacy and security, transparency, accountability and reinforcement of positive human values. This was followed by another approach document that articulated the way forward for operationalizing the above principles.[555]

This document advocated a risk-based approach whereby the extent of regulatory controls over an AI system should be commensurate with the potential for harm posed by it. It, however, shied away from suggesting any kind of concrete regulatory measures for AI governance. The Indian government has maintained a similar stand in other policy communications. In a recent Parliamentary discussion, the Minister for Electronics and Information Technology clarified that while there was a need to encourage the use of AI and promote best practices to minimize harms, the government had no plans of bringing about a law on AI regulation.[556]

[553] NITI Aayog. (2018, June). Discussion Paper: National Strategy for Artificial Intelligence. Retrieved from <https://indiaai.gov.in/documents/pdf/NationalStrategy-for-AI-Discussion-Paper.pdf>.

[554] NITI Aayog. (2021a, February). Responsible AI Approach Document for India: Part 1 – Principles for Responsible AI. Retrieved from <https://www.niti.gov.in/sites/default/files/2021-02/Responsible-AI-22022021.pdf>.

[555] NITI Aayog. (2021b, August). Responsible AI Approach Document for India: Part 2 – Operationalizing Principles for Responsible AI. Retrieved from <https://www.niti.gov.in/sites/default/files/2021-08/TowardsResponsibleAI-newReport3.pdf>.

[556] The Hindu (2023, April 5). No plan to regulate AI: IT Ministry tells Parliament. Retrieved from <https://www.thehindu.com/news/national/no-plan-to-regulate-ai-it-ministry-tells-parliament/article66702044.ece>.

In November 2022, the NITI Aayog published yet another discussion paper that evaluated the principles and governance frameworks articulated earlier specifically in the context of facial recognition technologies.[557] As a part of this exercise, it undertook a deeper dive into one particular project, called the Digi Yatra project, which is a facial recognition based system for entry and boarding at Indian airports. The NITI Aayog paper on Digi Yatra is accompanied by other policy documents about the project. Notably, an explanatory note issued by the Ministry of Civil Aviation (MoCA) in 2018[558] and press releases issued from time to time announcing its different stages of development and deployment. The privacy and other human rights implications of Digi Yatra,[559, 560] the need for transparency in such large scale public deployments of AI and the availability of a fair amount of information about this project make Digi Yatra a suitable case study for the present analysis.

Set against this background, Section 2 of this paper offers an overview of how the principle of transparency has been articulated in the AI governance discussions in India and in select international principles. Section 3 establishes how the need for transparency permeates through the lifecycle of an AI project. Specifically, the type of transparency expected from different actors may vary depending on their role and location in the value chain of the AI project. The paper identifies the policy layer, the technical layer and the operations layer as three key, and often overlapping, components of an AI system's value chain. It then applies this multi-layered expectation of transparency analysis to the Digi Yatra project. Section 4 summarizes the main findings and conclusions.

[557] NITI Aayog. (2022, November). Responsible AI Adopting the Framework: A Use Case Approach on Facial Recognition Technology. Retrieved from <https://niti.gov.in/sites/default/files/2023-03/Responsible-AI-AIForAll-Approach-Document-for-India-Part-Principles-for-Responsible-AI.pdf>.

[558] Ministry of Civil Aviation, Government of India (2018, August 9). Digi Yatra Biometric Boarding System: Reimagining Air Travel. Version 5.2. Retrieved from <https://www.civilaviation.gov.in/sites/default/files/Digi%20Yatra%20Policy%2009%20Aug%2018.pdf>.

[559] Jain A. (2022, January 18). The dangers of DigiYatra & facial recognition enabled paperless air travel. Internet Freedom Foundation. Retrieved from <https://internetfreedom.in/dangers-of-digiyatra/#:~:text=The%20scheme%20aims%20to%20make,any%20remedies%20or%20regulatory%20framework>.

[560] Kodali S. (2023, February 13)., How 'Digi Yatra' Can Potentially Be Used to Police Your Travel. The Wire. Retrieved from <https://thewire.in/rights/digi-yatra-privacy-biometric-travel>.

## 14.2 Unpacking the Principle of AI Transparency

The principle of AI transparency, as seen in various principles and recommendations, can be unpacked at two levels. The first is to understand *why* the principles call for a need for greater transparency and second, is there any guidance on *how* transparency is to be achieved.

The OECD's AI Principles[561] club transparency with explainability in order to foster a general understanding of the system being used and to enable people to challenge its outcomes (Principle IV, 1.3). The goal of transparency in this context includes facilitating an understanding of when someone is interacting with an AI system, being able to understand its outcomes and challenge the logic behind its functioning. Similarly, UNESCO's AI principles[562] link transparency with the efficient functioning of liability regimes and building scope for challenging the decisions and outcomes of AI systems. This reinforces the link between AI transparency, explainability, trustworthiness and accountability. Further, UNESCO's approach also draws a connection between the pursuit of transparency as a means for the effective functioning of democratic governance and enabling greater public scrutiny. In India, the principles for responsible AI describe transparency as a requirement for the design and functioning of AI systems to be amenable to external scrutiny and audit.[563] The objective being that the use of AI should be fair and honest and support accountability.

On the issue of how to achieve transparency, the OECD's AI Principles call upon AI actors to ensure transparency by providing meaningful information that is appropriate to the context, and consistent with the state of the art. These principles define 'AI actors' to mean all organizations and individuals that play an active role in the lifecycle of an AI system, including its deployment and operations. The goal of transparency also interacts with other requirements

[561] OECD (2019, May 21). Organisation for Economic Co-operation and Development Recommendations of the Council on Artificial Intelligence. Retrieved from <https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449>.

[562] UNESCO (2022). United Nations Educational, Scientific and Cultural Organization Recommendation on the Ethics of Artificial Intelligence. Retrieved from <https://unesdoc.unesco.org/ark:/48223/pf0000381137>, p. 22.

[563] NITI Aayog, 2021a, (op. cit.).

and objectives, such as that of trustworthiness and scrutiny. For instance, the UNESCO principles place specific emphasis on the transparency of ethical impact assessments, 'which should also be multidisciplinary, multi-stakeholder, multicultural, pluralistic and inclusive' in character.[564]

In India's case, although the responsible AI principles do not elaborate on how to operationalise transparency, the second approach paper by NITI Aayog offers more guidance in this regard. Notably, while calling for a risk-based approach to AI regulation, the document notes that the sociotechnical system as a whole needs to be considered while assessing the potential for harm from a particular project. The approach paper then lays out the expectations from different groups of stakeholders, such as the government, the private sector and research institutions.[565] The NITI Aayog has also placed significant emphasis on the need for transparency in the procurement processes followed by the government while selecting a technology vendor and in the error rates encountered in the course of implementation of the project.[566]

An annex to NITI Aayog's responsible AI principles contained examples of model transparency mechanisms.[567] The list included Google's Model Card Toolkit, Microsoft's datasheets for datasets, IBM's Fact Sheet project. All of these examples speak to the issue of transparency at the level of the model or the algorithm. This focus of algorithmic transparency is also reflected in other initiatives like the AI Algorithmic Transparency Tool developed by researchers at the Tokyo University with private sector stakeholders and published by the OECD.[568] However, as noted above, various AI principle documents do make it clear that the applicable principles should extend through the life cycle of the AI project covering a range of actors. As per India's approach paper on operationalizing responsible AI, this includes the different components of an algorithmic application as well as

[564] UNESCO, 2022 (op. cit.), p. 26.

[565] NITI Aayog, 2021b (op. cit.).

[566] NITI Aayog, 2022 (op. cit.).

[567] NITI Aayog, 2021a, (op. cit.), p. 47.

[568] Tonfi Y., Masayuk O., Hiraku M., Kirihito Y. and Yuta N. (2023, April 13). AI Algorithm Transparency Toolkit: A Proposal for a Governance System to Enable Society to Accept and Benefit from AI-based Innovations.

the actors involved in all stages from its design to implementation to evaluation.[569] Accordingly, the next section of the paper takes a more granular look at the transparency expectations of AI systems, beyond just algorithmic or model-related transparency, using a multi-layered approach.

## 14.3 Multi-layered Expectations of AI Transparency

The lifecycle of an AI system consists of many stages and the principles of AI governance may apply differently at each stage. The OECD's observatory of AI tools classifies the AI life cycle into five stages, namely, planning and design, collection and interpretation of data, building and interpreting the model, verification and validation, deployment, operation and monitoring.[570] The actors involved in each of these stages could vary depending on the nature of the sector and ownership model of the project. For instance, an AI project undertaken by a large technology company may have most of these steps taking place within the same organization. Or it may involve the outsourcing of certain specific functions, like data collection, to third party firms. In contrast, the large-scale deployment of an AI system in the public sector would typically involve a broader range of actors and a complex series of interactions among them. This can be demonstrated using the selected case study of the Digi Yatra project.

In August, 2018 India's MoCA announced its plans to launch a biometric system for airport entry and boarding management procedures under the Digi Yatra project (MoCA, 2018). The document set out the design and detailed process flow of the project. It also revealed that MoCA had already been working on this initiative for over a year with a Technical Working Committee constituted by it. Subsequently, a non profit company called the Digi Yatra Foundation was created in 2019 to give effect to the Digi Yatra Central Ecosystem.[571] The shareholding of this entity was held by the Airport Authority of India, a statutory authority under the MoCA and six other companies

[569] NITI Aayog, 2021b (op. cit.), p. 13.

[570] OECD (n.d.). Catalog of Tools & Metrics for Trustworthy AI. Retrieved from <https://oecd.ai/en/catalogue/tools>.

[571] NITI Aayog, 2022 (op. cit.), p. 27.

that operate as special purpose vehicles (SPVs) for the operation of airports in the cities of Cochin, Bengaluru, Delhi, Hyderabad and Mumbai.[572] The project was launched at three airports, New Delhi, Varanasi, and Bengaluru, in December, 2022 and has slowly been extended to a number of other cities.[573]

Alongside these operational developments, the NITI Aayog also became involved in the effectuation of the DigiYatra project. It collaborated with the Digi Yatra Foundation, the Atal Innovation Mission – a government program to encourage innovation and entrepreneurship – and Amazon Web Services for the selection of the technical implementation partner for DigiYatra. Pursuant to this, an entity known as Dataevolve Solution was selected to implement the technical specifications of Digi Yatra.[574] While the selection of Dataevolve Solution took place through an open challenge, the mode and specifics relating to the involvement of Amazon Web Services, which is reportedly partnering with Dataevolve,[575] remains less clear. In addition, each airport authority is selecting a technology implementation partner for giving effect to the facial recognition system on the ground.[576] The operation of the DigiYatra system requires individuals to download an app on their smartphones, which operates as a digital identity wallet to be used while accessing the airport for boarding a flight. Accordingly, platforms like Google Play Store and Apple App Store that host the DigiYatra app also become relevant stakeholders in the implementation of the system.

Besides the government think tank, NITI Aayog, certain private think tanks are known to have played a role in the design and analysis of the DigiYatra system. One of these is the Indian Software Industry Roundtable (iSPIRT), a prominent Indian software industry supported

[572] Digi Yatra Foundation (n.d.). About us. Retrieved from <https://digiyatrafoundation.com/>.

[573] Ministry of Civil Aviation, Government of India. (2023(a), August 11). Digi Yatra to be launched at 6 more airports. Press Information Bureau. Retrieved from<https://pib.gov.in/PressReleaseIframePage.aspx?PRID=1947913>.

[574] Ministry of Civil Aviation, Government of India. (2023(b), February 2). Digi Yatra to be implemented at Kolkata, Pune, Vijayawada and Hyderabad Airports by March 2023. Press Information Bureau. Retrieved from <https://pib.gov.in/PressReleaseIframePage.aspx?PRID=1895743>.

[575] Money Control. (2023, February 2). This Hyderabad-based startup is behind the airport walk-in app Digi Yatra. Retrieved from <https://www.moneycontrol.com/news/business/this-hyderabad-based-startup-is-behind-the-airport-walk-in-app-digiyatra-9989271.html>.

[576] The Print. (2023, June 20). IDEMIA Selected as Technology Partner by DIAL for DigiYatra. Retrieved from<https://theprint.in/ani-press-releases/idemia-selected-as-technology-partner-by-dial-for-digiyatra/1634106/>.

think tank that claims to have 'been intimately involved in' the Digi Yatra project.[577] The exact scope of iSPIRT's involvement in the conceptualisation and implementation of the project is, however, not clear from the official documentation on the project. The NITI Aayog's facial recognition approach paper recognises the role of another private entity, the legal think tank Vidhi Centre for Legal Policy, as its knowledge partner in developing the analysis of the Digi Yatra project based on the responsible AI principles.[578]

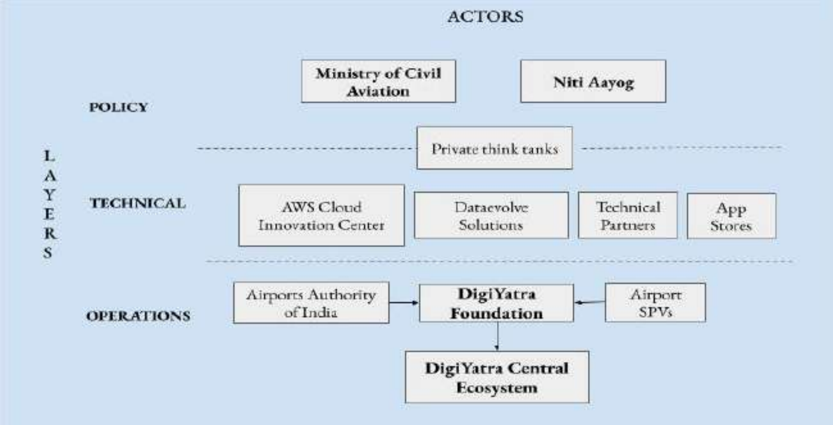


Figure 1: Actor map of India's Digi Yatra System.

Figure 1 depicts the three broad layers in the life cycle of the DigiYatra project – policy design, technical design and operations. These layers cannot be regarded to be mutually exclusive. Nor are they necessarily sequential in nature. For instance, the technical and operations layers operate in tandem for the implementation of the project, the lessons from which may continue to inform the policy design. The layer-wise framing is, however, useful for understanding the different categories of actors involved in implementing the different functions and in thinking about the nature of transparency that would be expected at each layer.

The preceding discussions demonstrate that both the MoCA and NITI Aayog displayed a certain degree of transparency in putting

[577] Singh, S. (2022, February 1). iSPIRT Foundation's Response to Union Budget 2022. iSPIRT Blog. Retrieved from <https://pn.ispirt.in/response-to-union-budget-2022/>.

[578] NITI Aayog, 2022 (op. cit.), p. iii.

out information about the Digi Yatra project. However, the nature of the transparency exercised by the MoCA was more in the nature of imparting information about the project to the public rather than a deliberative discussion on the need for the system, its design choices and risk factors. The NITI Aayog, on the other hand, did put up its approach paper on facial recognition, which included the analysis on DigiYatra, for public comments. However, this came at a stage when the project was already at an advanced stage of deployment.

At the technical design and implementation level, it has been documented that the selection of various technical partners, like Dataevolve Solutions and airport-specific partners took place through an open challenge and a tender process, respectively.[579] However, as discussed above, the technical partners involved in the project also included a number of other entities, like Amazon Web Services and the private think tank iSPIRT, the basis for whose selection to participate in the project has not been made known. This gap is also reflected in the NITI Aayog's analysis about the importance of transparency in procurement processes. The entire focus of this discussion is centered around the procurement of the technology itself without acknowledging the role of transparency in the procurement of ideas and advisory services surrounding AI adoption. As a result, private and ad hoc arrangements for the procurement of technical advice and services, which often accompany the adoption of automated systems in India, tend to remain outside the fold of AI transparency.

Finally, there is the important question of transparency at the operational level. One of the major levers for public transparency in India flows from the Right to Information Act, 2005 (RTI Act). This law creates a right for citizens to access information that is under the control of public authorities. However, the position expressed by the Indian government on Digi Yatra is that since the Digi Yatra Foundation is not a government body, but rather a non profit entity controlled by participating airports, it does not fall under the purview of the RTI Act (Sarasvati, 2023). The operational design of the Digi Yatra project, which is spearheaded and led, but not owned, by the government, therefore creates a serious roadblock to its transparency.

[579] DigiYatra Foundation E-Procurement (n.d.). Retrieved from <https://digiyatra.procure247.com/home>.

So far, the Digi Yatra Foundation has put out only some basic information, like its privacy policy and frequently asked questions, in the public domain. There is very little transparency about the day-to-day functioning of the system and the issues encountered in the process. This would include granular data about the adoption of the system, its success and failure rates and functioning of the redress mechanisms to deal with any difficulties encountered by individuals. The absence of timely and complete information about the technical and operational performance of the project presents a setback to the transparency and, by extension, explainability and accountability of the system.

## 14.4 Conclusion

This paper offered a broad overview of how the principle of transparency has been articulated in the AI governance discussions in India and in select international principles. While doing so it focused on the *why* and *how* of AI transparency obligations, as seen in the studied instruments. The paper argued that the need for transparency permeates through the lifecycle of an AI project and the type of transparency expected from different actors would vary depending on their role and location in the AI value chain.

Using the example of India's Digi Yatra project, the paper identified the policy layer, the technical layer and the operations layer as three key, often overlapping, components in an AI system's value chain. It studied the nature of transparency displayed by key actors, like the MoCA, the NITI Aayog and the Digi Yatra Foundation, in the functioning of this system. The observed transparancy was mainly in the form of putting out information about the project in the public domain and in the procurement of technical implementation partners.

Further, the paper highlighted three main gaps in the discharge of transparency obligations across the system's layers. First, it was observed that the policy transparency surrounding the Digi Yatra project mainly served the purpose of imparting information to the public rather than meaningful deliberations about its necessity and design. Second, the paper pointed to the existence of certain private and ad hoc arrangements in the procurement of technical and policy advisory services relating to the project. Such arrangements are as intrinsic to the design and outcomes of an AI project as the procurement of the project's technological components and must, therefore, be viewed with a similar sense of urgency in terms of ensuring transparency. Third, the paper emphasized a gap in the project's operational transparency caused by the fact that the Digi Yatra Foundation was not being treated as a public authority under the right to information framework. Its treatment as such would have compelled the Digi Yatra Foundation to provide more granular data about the day-to-day functioning of the system, on a suo moto basis as well as upon public request.[580]

The multi-layered transparency analysis suggested here can be useful in unveiling issues of transparency and accountability across the policy, technical and operational layers of any AI system. Absent such an approach, a large part of the focus of AI transparency conversation may remain on algorithmic or technical transparency while ignoring the procedural and administrative elements. This becomes particularly relevant for large, public-facing, AI systems that involve a complex set of actors, with differential transparency expectations from the system's participants.

[580] Saravati N.T. (2023, March 3). India's Civil Aviation Ministry Says Information On Digi Yatra Cannot Be Sought Under Right To Information. Medianama. Retrieved from <https://www.medianama.com/2023/03/223-civil-aviation-ministry-digi-yatra-right-to-information/>.

# 15 Principles for Enabling Responsible AI Innovations in India: An Ecosystem Approach

Mr Kamesh Shekhar, Programme Manager, Data Governance and privacy, The Dialogue;

Ms. Jameela Sahiba, Senior Programme Manager, Emerging Tech, The Dialogue;

Mr. Bhavya Birla, Research Associate, The Dialogue.

## Abstract

With the rapid proliferation of artificial intelligence (AI) across various domains, discussions surrounding responsible AI have become ubiquitous. These versatile technologies are transforming the nature of our work, interactions, and lifestyles. We are on the brink of witnessing a transformational shift comparable to the impact of the printing press, which revolutionised the world six centuries ago. Within this transformative landscape, our research paper assumes extraordinary significance. The paper proposes a principle-based multistakeholder approach which resonates with the foundational values of responsible AI envisioned by various jurisdictions geared towards ensuring that AI innovations align with societal values and priorities. Currently, there are various kinds of literature on the risk management of AI at the development level focusing on uni-stakeholder, i.e., AI developers. In a rapidly changing landscape, regulatory interventions must withstand the test of time. This is the primary reason why draft regulations in development or in the process of becoming a law must be principle-based. The approach to this paper for establishing an effective governance structure for AI would involve multi-stakeholders, including AI developers, AI deployers and impact population, where we map principles for different stakeholders within the AI ecosystem to make it trustworthy and safe. This paper, through a meta-analytic literature review, will also effectively contribute toward the discussion on developing an effective governance structure for AI to enhance its opportunities while mitigating its impact and harms at the international level, where the importance of global coordination and cooperation has become predominant now more than ever.

***For user readability and holistic contextualising of our paper, we believe it is imperative for us to explain key definitions of terms used across our paper in layman's terms:***

**1. AI Ecosystem:** AI Ecosystem refers to the interconnected environment of organisations, individuals and governments involved in the development, deployment and use of AI systems.

**2. AI System:** An AI system is an AI-powered, machine-based system that is capable of influencing the environment by producing an output (predictions, recommendations or decisions) for a given set of objectives.

**3. AI Lifecycle:** An AI life cycle refers to the sequential stages involved in the development, deployment and use of AI systems. The AI lifecycle consists primarily of the following stages: i) design, data and models; ii) verification and validation; iii) deployment; and iv) operation and monitoring.

**4. AI Actors:** AI actors are those who play an active role in the AI system lifecycle, including organisations and individuals that deploy or operate AI.

**5. AI Developer:** A natural person or legal entity (within both the public and private sectors) who develop AI systems for market consumption while they may not necessarily deploy and use the same technology.

**6. AI Deployer:** A natural person or legal entity (within both the public and private sectors) who procure, employs, deploys and operates AI systems not necessarily developed by themselves.

**7. Impact Population:** A natural person who directly or indirectly uses, engages, and is impacted or affected by the AI systems.

**8. Impact:** Impacts arise when the responsible parties or AI actors acknowledge, explain or take actions to mitigate the harms.

**9. Harms:** Harms refer to the negative or detrimental outcomes of AI systems on the end-users.

**10. Responsible AI:** The concept of responsible AI recognizes the need to ensure safe, beneficial, ethical and fair use of AI technologies to ensure societal progress, economic growth, and sustainable development of technology.

## 15.1 Introduction

The internet is advancing at an exponential pace; where within a short period, we have seen a transformation of two-dimensional Web 2.0 to technological developments like Artificial Intelligence, which senses the ethos to offer responses to our queries which is almost near to human reply. Artificial Intelligence is one of the driving forces of change that will shape the Internet in the coming days.[581] Therefore, making Artificial Intelligence trustworthy will contribute to making the Internet trustworthy.[582] For instance, as the Internet evolved, the face of Web 2.0 has been the intermediaries like social media platforms, search engines etc.,[583] which has brought to the floor the importance of the safe harbour and online safety debates; similarly, with the evolution to Web 3.0, increasingly we see that Artificial Intelligence is becoming the face of Internet. Therefore, to exert individuals' trust in the internet, tackling concerns emerging with Artificial Intelligence is important.

This paper will effectively contribute toward the discussion on developing an effective governance structure for AI to enhance its opportunities while mitigating its impact and harms. There are various kinds of literature on the risk management of AI at the development level focusing on uni-stakeholder, i.e., AI developers.[584] However, the approach to this paper for establishing an effective governance structure for AI would involve multi-stakeholders, including AI developers, AI deployers and impact population, where we map principles for different stakeholders within the AI ecosystem to make it trustworthy and safe.

[581] Thomas, M. (2022, August 9). *The future of AI: How artificial intelligence will change the world*. Built-In. Retrieved June 20, 2023, from <https://builtin.com/artificial-intelligence/artificial-intelligence-future>.

[582] While typical use-cases of AI technologies is beyond traditional experience of using internet, however as rightly identified by the Internet Society's Global Internet Report 2017, Artificial Intelligence is one of the driving forces of change that will shape the Internet in the coming days.

[583] O'Neill, S. (2022, January 7). *What's The Difference Between Web 1.0, Web 2.0, And Web 3.0?* MarTech Alliance. Retrieved August 16, 2023, from <https://www.lxahub.com/stories/whats-the-difference-between-web-1.0-web-2.0-and-web-3.0>.

[584] Rogers, J. (2023, January 11). *Artificial intelligence risk & governance. AI & Analytics for Business*. Retrieved June 20, 2023, from <https://aiab.wharton.upenn.edu/research/artificial-intelligence-risk-governance/>.

Chapter 2 of the paper will discuss various global developments in regulating Artificial Intelligence and operationalising key principles to set the context. Following this, in Chapter 3, we will list the five critical implications of AI solutions namely, exclusion, false predictions, copyright infringement, privacy infringement, and information disorder and try to map out the extent to which AI developers, AI deployers, and the impact population contribute towards manifesting the same. In addition, in Chapter 3, we propose a principle-based multistakeholder approach where we map the principles to be followed by stakeholders, namely AI developers, AI deployers and impact population at appropriate stages. Chapter 3 also discusses indicative operationalisation strategies for AI developers, AI deployers, and the impact population to imbibe the mapped principles. Finally, Chapter 4 discusses the domestic government's role in implementing the principle-based multistakeholder approach.

## 15.2 Status-quo of AI Regulations

Regulatory developments have cropped up worldwide to enhance AI risk management and trustworthiness in the recent past (namely, NITI Aayog's National Strategy for Artificial Intelligence,[585] OECD AI principles,[586] G20 AI Principles,[587] Australia's AI Intelligence Ethics Framework and AI Ethics Principles,[588] EU Ethics Guidelines for Trustworthy AI,[589] EU-US TTC Joint Roadmap for Trustworthy AI and Risk Management,[590] NIST's AI Risk Management Framework,[591] Germany, Artificial Intelligence Strategy 2018,[592] Singapore National

[585] NITI Aayog. (June 2018). *National Strategy for Artificial Intelligence #AIforAll. (2018)*. Niti Aayog. <https://niti.gov.in/sites/default/files/2019-01/NationalStrategy-for-AI-Discussion-Paper.pdf>.

[586] OECD. (2019). *Recommendation of the Council on Artificial Intelligence.* OECD Legal Instruments. <https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449>.

[587] G20. (2019). *G20 AI Principles.* <https://www.mofa.go.jp/policy/economy/g20_summit/osaka19/pdf/documents/en/annex_08.pdf>.

[588] Australian Government. (2019). *Australia's AI Ethics Principles*. <https://www.industry.gov.au/publications/australias-artificial-intelligence-ethics-framework/australias-ai-ethics-principles>.

[589] European Commission. (2019). *Ethics Guidelines for Trustworthy AI.* High-Level Expert Group on Artificial Intelligence, <https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=60419>.

[590] European Commission, (2022) *TTC Joint Roadmap on Evaluation and Measurement Tools for Trustworthy AI and Risk Management*. <https://ec.europa.eu/newsroom/dae/redirection/document/92123>.

[591] National Institute of Standards and Technology. (2023, January). *Artificial Intelligence Risk Management Framework.* NIST Technical Series Publications. <https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf>.

[592] German Federal Government. (2020, December). *National AI Strategy.* KI Strategie. <https://www.ki-strategie-deutschland.de/files/downloads/Fortschreibung_KI-Strategie_engl.pdf>.

AI Strategy 2019,[593] USA's National Artificial Intelligence Research and Development Strategic Plan 2023,[594] France's AI for Humanity 2017,[595] European Union's Artificial Intelligence for Europe 2018,[596] European Union's The Artificial Intelligence Act, 2023,[597] United Kingdom's A Pro-Innovation Approach to AI Regulation 2023,[598] Japan's Social Principles of Human-Centric AI 2019,[599] The Global Partnership on Artificial Intelligence's AI principles,[600] United Nations' Principles for Ethical Use of AI in UN 2022,[601] UNESCO Ethics of Artificial Intelligence,[602] and other private sector frameworks.[603] Against this backdrop, this chapter will discuss various global developments in regulating Artificial Intelligence and operationalising key principles. While various developments are happening around regulating AI worldwide, this chapter discusses some of the critical frameworks that have emerged at the lateral and multilateral levels across the globe.

An analysis of pathways taken by some of the critical jurisdictions on regulating AI shows that the ounce of tackling concerns about AI is overtly on AI developers. This paper will try to address the gap through

[593] Smart Nation Digital Government Office. (2019, November). *National Artificial Intelligence Strategy*. Smart Nation Singapore. <https://www.smartnation.gov.sg/files/publications/national-ai-strategy.pdf>.

[594] National Science and Technology Council. (2023, May). *The National Artificial Intelligence R&D Strategic Plan 2023 Update*. The White House. <https://www.whitehouse.gov/wp-content/uploads/2023/05/National-Artificial-Intelligence-Research-and-Development-Strategic-Plan-2023-Update.pdf>.

[595] Villani, C. (2018, March). *For A Meaningful Artificial Intelligence: French Strategy*. AI for humanity. <https://www.aiforhumanity.fr/pdfs/MissionVillani_Report_ENG-VF.pdf>.

[596] European Commission. (2018, April). *Artificial Intelligence for Europe*. EUR-Lex — Access to European Union law. <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=COM:2018:237:FIN>.

[597] European Commission. (2021, September). *The Artificial Intelligence Act*. The AI Act. <https://artificialintelligenceact.eu/the-act/>.

[598] Department for Science, Innovation and Technology. (2023, March). *A pro-innovation approach to AI regulation*. <https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1146542/a_pro-innovation_approach_to_AI_regulation.pdf>.

[599] The Government of Japan. (2019, February). *Social Principles of Human-Centric AI*. <https://www.cas.go.jp/jp/seisaku/jinkouchinou/pdf/humancentricai.pdf>.

[600] The Global Partnership on Artificial Intelligence's AI principles. (2020, June). *Global Partnership on Artificial Intelligence* – GPAI. <https://gpai.ai/about/>.

[601] UN System Chief Executives Board for Coordination. (2022, September). *Principles for the Ethical Use of Artificial Intelligence in the United Nations System*. United Nations – CEB. <https://unsceb.org/sites/default/files/2022-09/Principles%20for%20the%20Ethical%20Use%20of%20AI%20in%20the%20UN%20System_1.pdf>.

[602] UNESCO. (2023, April 20). *UNESCO adopts first global standard on the ethics of artificial intelligence*. <https://www.unesco.org/en/articles/unesco-adopts-first-global-standard-ethics-artificial-intelligence>.

[603] Schiff J, D., Borenstein, J., & Laas, K. (2021, April 12). *AI ethics in the public, private, and NGO sectors: A review of a global document collection*. Montreal AI Ethics Institute. <https://montrealethics.ai/ai-ethics-in-the-public-private-and-ngo-sectors-a-review-of-a-global-document-collection/>.

discussion at the ecosystem level. This analysis also showcases that there is a lot of effort and literature on risk management of AI at the development level focusing on uni-stakeholders, i.e., AI developers.[604] However, these fall through the cracks as we leave other players undiscussed. Therefore, in chapter three, we will discuss establishing an effective governance structures for AI regulation at a domestic and inter-governmental level involving multistakeholders, i.e., AI developers, AI deployers and impact population, where various principles will be mapped to different stakeholders towards making AI trustworthy and safe.

## 15.3 Principle-based Multi-Stakeholder Approach – An Ecosystem-Level Intervention

It is crucial to minimise the impact and harms of Artificial Intelligence to make it a success. As discussed in the previous chapter, countries across the globe are taking steps to regulate AI, such as the recent draft of Brazil's AI Bill, the EU's AI Bill, and the US National Institute of Standards and Technology's AI RMF, NITI Aayog's responsible AI principles. While these regulatory measures are trying to make AI systems trustworthy through risk management, there is less discussion on how we can tackle the adverse implications of AI artificial intelligence at the ecosystem level, involving other stakeholders like AI deployers and the impact population. Besides, in a rapidly changing landscape, regulatory interventions must withstand the test of time. This is the primary reason why draft regulations in development or in the process of becoming a law must be principle-based.[605]

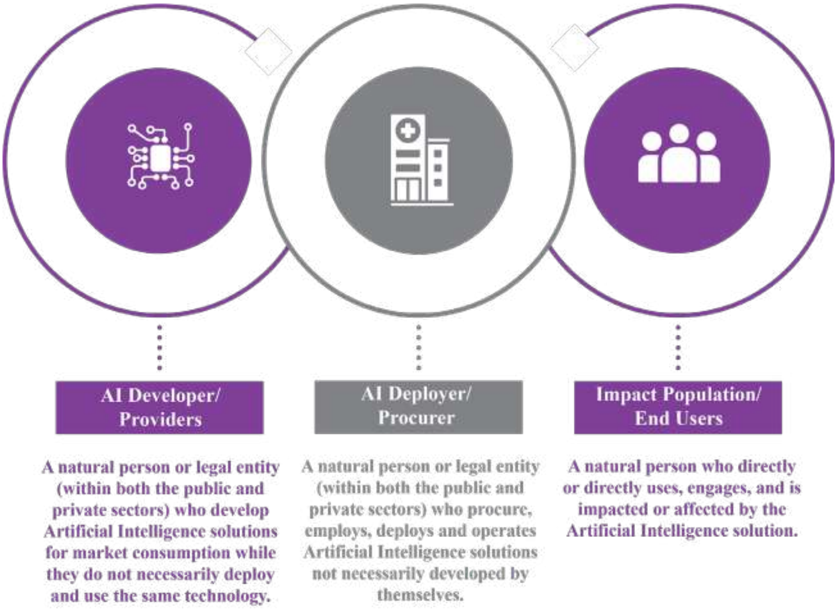
Therefore, through this chapter, we suggest a principle-based multi-stakeholder approach where we discuss various principles across the AI lifecycle bucketed and mapped to respective stakeholders within the AI ecosystem.[606]

[604] Rogers, J. (2023, January 11). *Artificial intelligence risk & governance. AI & Analytics for Business*. Retrieved June 20, 2023, from <https://aiab.wharton.upenn.edu/research/artificial-intelligence-risk-governance/>.

[605] Maithon, R. (2023, April 11). *India needs a principles-based approach to regulating AI*. Bharat Times. Retrieved June 20, 2023, from <https://news.bharattimes.co.in/india-needs-a-principles-based-approach-to-regulating-ai/>.

[606] The principles should be understood in their cultural, linguistic, geographic, and organizational context, and some themes will be more relevant to a particular context and audience than others. For instance, the definition of transparency or explainability in Brazil may not be the same concept in the US.

While there are various stakeholders within the AI ecosystem, this chapter covers the three key players, i.e., AI developers, AI deployers, and Impact Population. For the purpose of this chapter, three key stakeholders are defined as the following.[607]



AI Developer/ Providers

A natural person or legal entity (within both the public and private sectors) who develop Artificial Intelligence solutions for market consumption while they do not necessarily deploy and use the same technology.

AI Deployer/ Procurer

A natural person or legal entity (within both the public and private sectors) who procure, employs, deploys and operates Artificial Intelligence solutions not necessarily developed by themselves.

Impact Population/ End Users

A natural person who directly or directly uses, engages, and is impacted or affected by the Artificial Intelligence solution.

Figure 1: Stakeholders.

The critical principles mapped for the above-discussed stakeholders in this chapter are advised by the frameworks developed by various governments, intergovernmental organisations, academia, civil society etc., in India and globally. Besides, the principles discussed in this chapter are the key universal and internationally recognised AI design and deployment principles embedded in various responsible AI frameworks across jurisdictions,[608] especially India.[609]

[607] The AI developer and AI deployers are not watertight compartments, whereas there are instances where the AI provider/developer could also be an AI operator/user. At such conditions, the entity or natural person must follow the principles bucketed for AI developers and AI deployers at different stages of the AI lifecycle.

[608] Shankar, V., & Casovan, A. (2022, May). *A framework to navigate the emerging regulatory landscape for AI.* The OECD Artificial Intelligence Policy Observatory – OECD.AI. <https://oecd.ai/en/wonk/emerging-regulatory-landscape-ai>.

[609] NITI Aayog. (2022, November). *RESPONSIBLE AI #AIFORALL Adopting the Framework: A Use Case Approach on Facial Recognition Technology.* | NITI Aayog. <https://www.niti.gov.in/sites/default/files/2022-11/Ai_for_All_2022_02112022_0.pdf>.

## 15.4 Mapping Harms and Impact across the AI Lifecycle

While we interchangeably use terms such as Impacts and Harms, they are technically not identical. The impacts can be defined as evaluative constructs used to gauge the socio-material harms[610] that can result from AI systems systematically and objectively.[611] These measurable outcomes allow us to understand the consequences of the interaction between AI technologies and individuals and society. For instance, the error rates of the AI solution, like the rate of inaccurate information, wrong predictions or disparate errors etc. Defining and measuring impacts allows us to understand the intended and unintended risks, benefits and harms that may arise when the procured AI deployers employ the AI solutions.

However, though the developed AI solutions are working as designed, adverse implications still crop out. This is where the other end of the puzzle, which is less discussed, comes into the picture, i.e., how AI deployers utilise the procured AI solutions for critical functions causing tangible and intangible harms.[612] For instance, as discussed above, the AI solutions might be producing an error or may be designed to capture some biased parameters to produce the suggested outcome; however, real-life harms of such outcomes only translate into action when AI deployers blindly use the same for making real-life decisions.[613]

Therefore, this shows that the distinction between harm and impact is rooted in the accountability and responsibility relationship among the stakeholders involved in the AI lifecycle, where both AI developers and AI deployers must follow some key principles to ensure adverse implications of AI solutions are

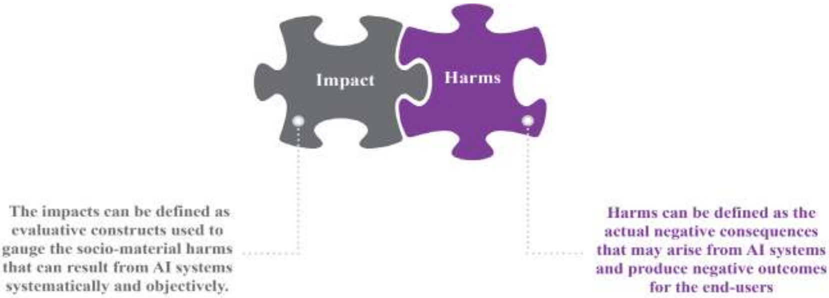
[610] Socio-Material harms in this context refer to the harms that a faulty operationalisation of AI technologies can have on end-users. These range from impacts emmanating from being subject to algorithmic decision making, AI powered bias and discrimination and even an invasion of privacy amongst others addressed at length under chapter three.

[611] Metcalf J, Moss E, Watkins E, Singh R, and Elish M. (2021, March). *Algorithmic Impact Assessments and Accountability: The Co-construction of Impacts.* ACM Digital Library. <https://dl.acm.org/doi/pdf/10.1145/3442188.3445935>.

[612] Horowitz, A., & Selbst, A. (2022, June). *The fallacy of AI functionality.* ACM Digital Library. <https://dl.acm.org/doi/fullHtml/10.1145/3531146.3533158>.

[613] It is important to note that the AI developer and AI deployers are not watertight compartments, whereas there are instances where the AI provider/developer could also be an AI operator/user. At such conditions, the entity or natural person must follow the principles bucketed for AI developers and AI deployers at different stages of the AI lifecycle.

tackled appropriately.[614] Besides, with the evolution of artificial intelligence into Generative AI solutions, real-life harms could also be caused by the impact population. For instance, when an AI solution produces baseless and misleading information, this starts a chain reaction of misinformation, which becomes a wild forest fire as unsuspecting impact populations start sharing the same misleading information within their own network.[615]



The impacts can be defined as evaluative constructs used to gauge the socio-material harms that can result from AI systems systematically and objectively.

Harms can be defined as the actual negative consequences that may arise from AI systems and produce negative outcomes for the end-users

Figure 2: Impacts vs. Harms.

While there are various forms of adverse implications emerging out of AI solutions, for the purpose of this section, we will be concentrating on five critical implications of AI solutions, i.e., exclusion, false predictions, copyright infringement, privacy infringement, and information disorder. The rationale behind choosing these critical implications is based on the cluster of cases reported on the same, which has been slightly higher.[616] The below illustration showcases how AI developers, AI deployers, and the impact population contribute towards orchestrating the five critical implications. In doing so, the illustration will also showcase at what stages within the AI lifecycle[617] "impact" and "harm" emerge

[614] Ryan, M. (2020, June 9). *Artificial intelligence ethics guidelines for developers and users: Clarifying their content and normative implications.* Discover Journals, Books & Case Studies | Emerald Insight. Retrieved June 20, 2023, from <https://www.emerald.com/insight/content/doi/10.1108/JICES-12-2019-0138/full/html>.

[615] Discussed in detail below.

[616] European Commission. (2020, March). *The ethics of artificial intelligence: Issues and initiatives.* European Parliament. <https://www.europarl.europa.eu/RegData/etudes/STUD/2020/634452/EPRS_STU(2020)634452_EN.pdf>; *Crime in India 2021 | National Crime Records Bureau.* (n.d.). **राष्ट्रीय अपराध रकिॉर्ड ब्यूरो**. Retrieved August 18, 2023, from <https://ncrb.gov.in/en/Crime-in-India-2021>.

[617] Advised by OECD and NIST AI lifecycle, however, slightly improvised to fit the model suggested in this paper.

and how AI developers, AI deployers, and impact populations are associated with the same.

While various forms of impact and harm could potentially contribute towards causing the identified adverse implication, for the purpose of this paper, we have mapped some of the predominant causes based on our meta-analytic literature review. Besides, the mapped causes in the form of impact and harm don’t exist in water-tight compartments, where some of them could apply universally and could be true for other adverse implications than the one they are mapped to.

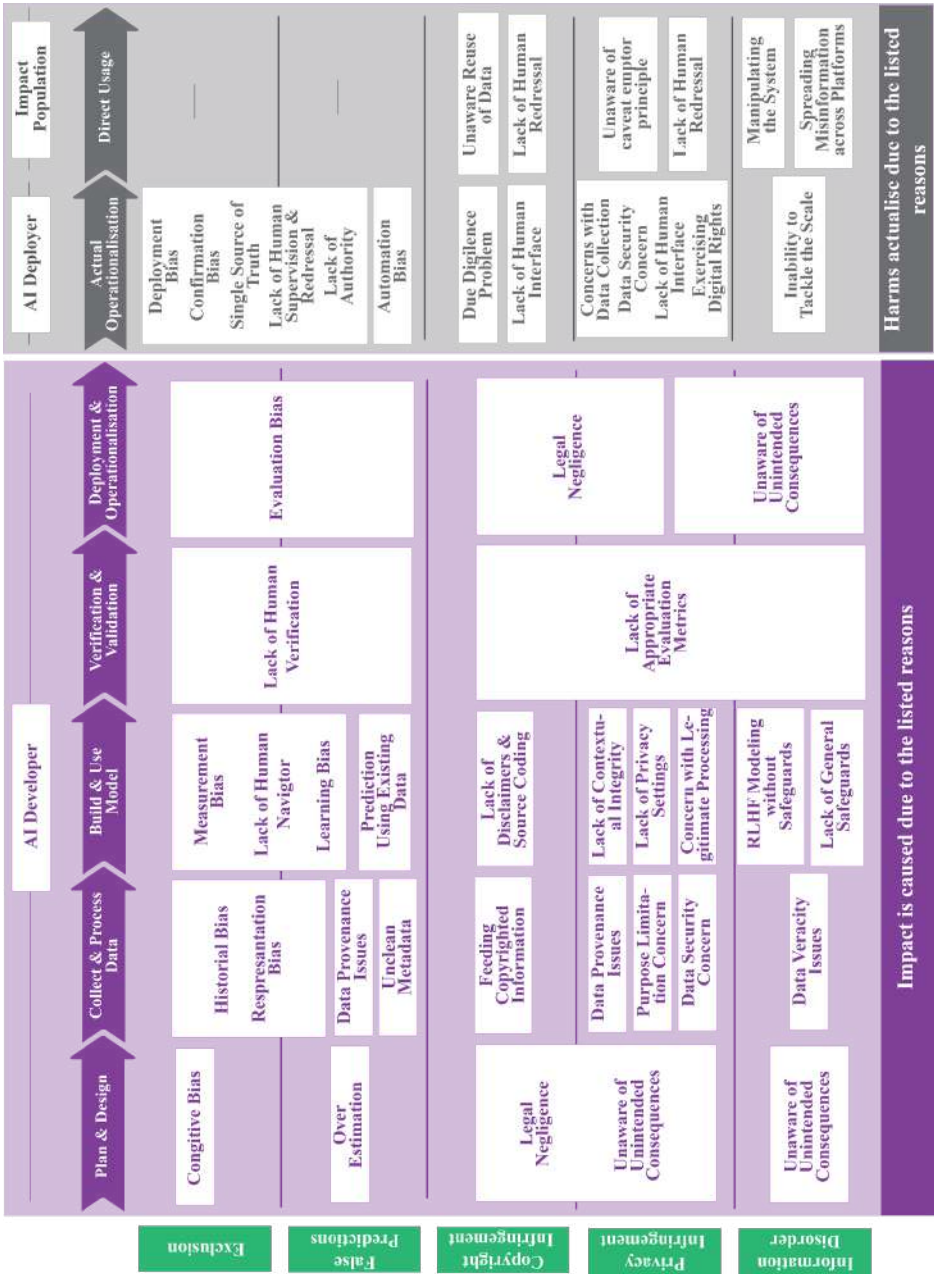

AI Developer
AI Deployer
Impact Population
Plan & Design
Collect & Process Data
Build & Use Model
Verification & Validation
Deployment & Operationalisation
Actual Operationalisation
Direct Usage
Exclusion
False Predictions
Copyright Infringement
Privacy Infringement
Information Disorder
Congitive Bias
Over Estimation
Historial Bias
Respresantation Bias
Data Provenance Issues
Unclean Metadata
Measurement Bias
Lack of Human Navigtor
Learning Bias
Prediction Using Existing Data
Lack of Human Verification
Evaluation Bias
Deployment Bias
Confirmation Bias
Single Source of Truth
Lack of Human Supervision & Redressal
Lack of Authority
Automation Bias
Legal Negligence
Unaware of Unintended Consequences
Feeding Copyrighted Information
Lack of Disclaimers & Source Coding
Lack of Appropriate Evaluation Metrics
Legal Negligence
Due Digilence Problem
Lack of Human Interface
Unaware Reuse of Data
Lack of Human Redressal
Data Provenance Issues
Purpose Limitation Concern
Data Security Concern
Lack of Contextual Integrity
Lack of Privacy Settings
Concern with Legitimate Processing
Concerns with Data Collection
Data Security Concern
Lack of Human Interface
Exercising Digital Rights
Unaware of caveat emptor principle
Lack of Human Redressal
Unaware of Unintended Consequences
Data Veracity Issues
RLHF Modeling without Safeguards
Lack of General Safeguards
Unaware of Unintended Consequences
Inability to Tackle the Scale
Manipulating the System
Spreading Misinformation across Platforms
Impact is caused due to the listed reasons
Harms actualise due to the listed reasons

Figure 3: Mapping Impact and Harms Across AI Lifecycle.

### 15.4.1 Exclusion

One of the main concerns around Artificial Intelligence is producing biased outputs, which could ultimately lead to the exclusion of impact populations traditionally excluded in real life. For instance, alternate credit lending platforms, which use the data points like education attainment, employment history, social media data etc., for underwriting and pricing loans, have been reported to discriminate against individuals based on historical biases.[618] Where individuals who attended colleges/ universities dedicated to historically vulnerable populations have been quoted a higher interest rate and were denied credit.[619]

India is a diverse and complex country with various historic dispositions like patriarchy, caste discrimination. Against this backdrop, one of the main concerns around AI is producing biased outputs. While AI solutions are not intentionally harmful, they replicate biases due to the biases present in its training data set and the way the algorithms are designed. Therefore, it is concerning when there is less clarity on the integrity, quality, and diversity of the data used for training the algorithms of these AI solutions. Besides, as these AI solutions are mostly predictive tools, they might unintendedly replicate the historic disposition causing discrimination and disproportionate harm to the vulnerable population. Moreover, the potential danger caused by AI is not just at the development stage but also at the deployment level, where harm could be caused by AI deployers who may abuse and misuse the technology.

### 15.4.2 False Predictions

Using an AI-based predictive tool can replicate bias due to the biases in its training set. For instance, AI technologies used for law enforcement purposes have been reported to bring out historical biases where for instance, systems have primarily assigned police parol to the neighbourhoods where discriminated populations reside. The incorrect crime predictions also feed into the system,

[618] Klein, A. (2022, March 9). *Reducing bias in AI-based financial services*. Brookings. Retrieved June 20, 2023, from <https://www.brookings.edu/research/reducing-bias-in-ai-based-financial-services/>.

[619] Klein, A. (2022, March 8). *Credit denial in the age of AI.* Brookings. Retrieved June 20, 2023, from <https://www.brookings.edu/research/credit-denial-in-the-age-of-ai/>.

creating a vicious cycle.[620] Similarly, the utilisation of AI in hiring tools used by companies and recruitment firms has been observed to increasingly discriminate against women. For instance, a company using AI solutions to hire a candidate for a particular role based on human-assigned ratings is reported to predict women as less suitable candidates than men, though the work profiles and qualifications of female candidates were at par with the male candidates. This false prediction scenario may be fed through historical bias against data recording the career growth trajectories of women across corporate settings.[621]

As discussed in Section 3.1.1 in the Indian context, the presence of the historically biased disposition against certain groups could aggravate adverse implications of the AI systems, like false predictions. While false predictions are one half of the story creating impact, the second half is when the AI deployers use those false predictions daily for determining eligibility, profiling etc., causing entry barriers, discrimination etc.

### 15.4.3 Copyright Infringement

A problem that could have legal repercussions enforced through monetary claims is that of an AI system infringing intellectual property rights.[622] Since some of the AI innovations, like generative AI technologies, are trained on a wide variety of language models, which include data such as books, articles, and journals, the output to be produced might have the risk of infringing on copyright texts leading to a violation of one's intellectual property rights. For instance, the outcome of generative AI solutions doesn't necessarily show original sources that it has used for deriving an answer; this might

[620] Sachoulidou, A. (2023, February 22). *Going beyond the "common suspects": To be presumed innocent in the era of algorithms, big data and artificial intelligence – artificial intelligence and law*. SpringerLink. Retrieved June 20, 2023, from <https://link.springer.com/article/10.1007/s10506-023-09347-w>.

[621] Goodman, R. (2023, February 27). *Why Amazon's automated hiring tool discriminated against women* | ACLU. American Civil Liberties Union. Retrieved June 20, 2023, from <https://www.aclu.org/news/womens-rights/why-amazons-automated-hiring-tool-discriminated-against>.

[622] For instance, Under the Indian IPR laws (Copyright Act 1957, Indian Patents Act, 1960 etc.) Patent and Copyright holders may sue AI developers for using their protected material for training foundation models. This has been observed across jurisdictions with prominent cases such as *Clakrson Law Firm v Open AI Case 3:23-cv-03199* in the United States of America, accessible from <https://clarksonlawfirm.com/wp-content/uploads/2023/06/0001.-2023.06.28-OpenAI-Complaint.pdf>.

cause an infringement of intellectual property. Besides, there is less clarity on the compensation mechanism for using the original work produced through human creativity. Some of the causes for copyright infringement are as follows.

### 15.4.4 Privacy Infringement

The AI solutions are trained using a massive amount of data to provide a human-like response. However, there is less clarity on the amount of personal information used by the AI developers as part of the training set and data protection measures taken to secure the same. Besides, there are also data security concerns where it is likely that AI solutions could expose confidential information of individuals causing identity theft, fraud etc. For instance, recently, Samsung spotted a generative AI solution leaking its confidential information as one of its unaware employees accidentally disclosed sensitive information while interacting with a generative AI solution.[623]

### 15.4.5 Information Disorder

While quick and easy access to information is useful, lack of understanding about the accuracy of the information received through AI solutions, especially with consumer-facing AI solutions like generative AI, is problematic – especially for high stake information like election-related information, health-related information etc. – given that disinformation and misinformation spread faster than the truth.

## 15.5 Mapping Principles for Stakeholders Across the AI Lifecycle

The various stakeholders within the AI ecosystem contribute in their capacities towards operationalising adverse implications, as discussed in Section 3.1. Therefore, to make the AI ecosystem safe, inclusive, and useful, it is essential to have a concerted effort at the ecosystem level where various stakeholders follow different principles at different stages of the AI lifecycle.

[623] Sharma, D. (2023, May 2). *Samsung restricts use of generative AI tools after employees leak sensitive data using ChatGPT*. India Today. Retrieved June 20, 2023, from <https://www.indiatoday.in/technology/news/story/samsung-restricts-use-of-generative-ai-tools-after-employees-leak-sensitive-data-using-chatgpt-2367448-2023-05-02>.

While these frameworks discuss principles for the responsible development of AI solutions, if the users misuse it and the impact population is unaware, it falls through the cracks. Therefore, we need a principle-based intervention that maps responsibilities and principles for various stakeholders within the AI ecosystem. While in the previous section, we did an implication-by-implication causation analysis, in this section, we will discuss the principles at the consolidated level mapped to various stakeholders to be followed at different stages, as illustrated below.

Collectively, we believe the mapped principles will enhance the trust of the impact population such that they feel at ease and safe using AI solutions.

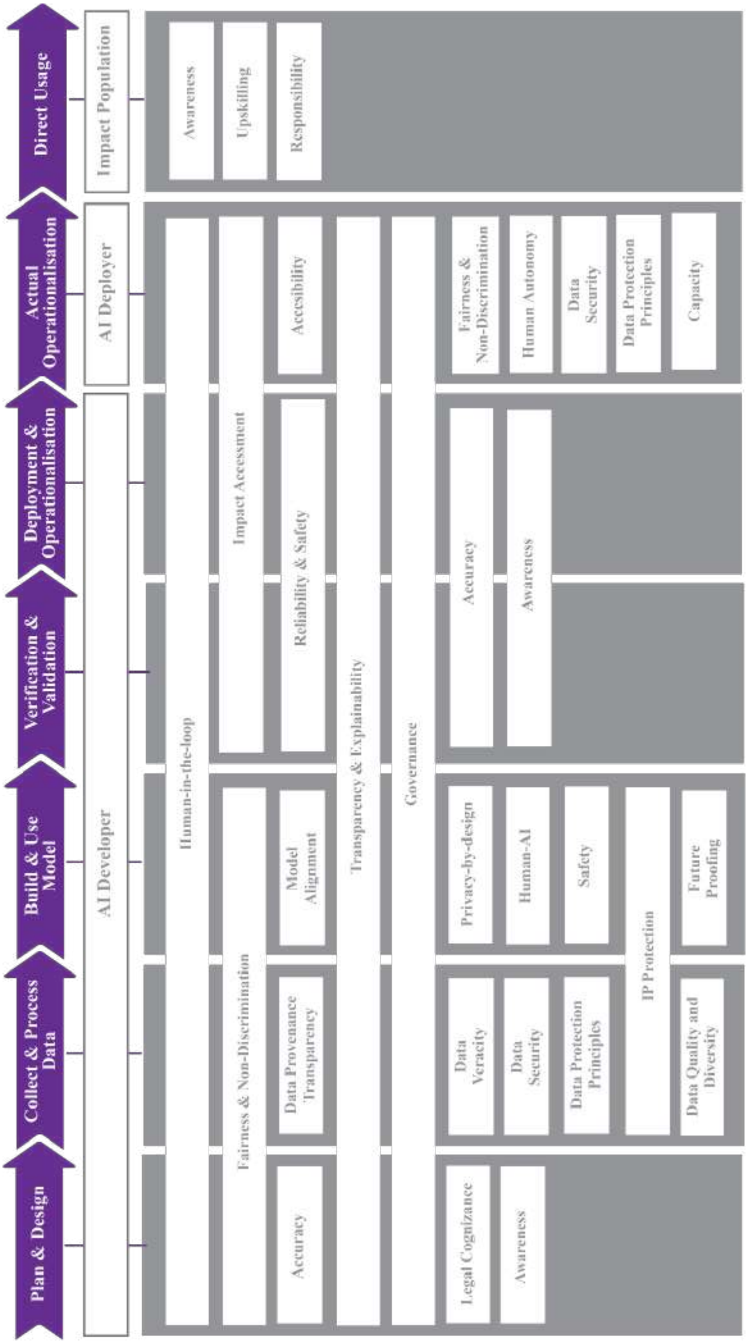


Figure 4: Mapping Principles for Stakeholders Across the AI Lifecycle.

## 15.6 Operationalisation of Principles by Various Stakeholders

To ensure the realisation of responsible AI, it is crucial to translate the principles discussed in the above chapter into tangible requirements. While there is a broad consensus regarding the core principles of responsible/ethical AI, there remains a lack of consensus on applying and implementing these principles within organisations effectively.

Besides, most of the AI principles' operationalisation frameworks have been at the level of risk management with less attention to the responsibilities, which lie at the level of AI deployers and Impact Population. Therefore moving from the uni-stakeholder approach, in this section, we will provide stakeholder-by-stakeholder strategies and means to operationalise the principles discussed in the previous section and their outcomes. While every principle would require/worth a separate research study in terms of operationalisation; however, the purpose of this paper is to map the principles and levers for operationalisation to a limited extent such that future research can be initiated on the same. We believe responsible AI can be effectively achieved by establishing concrete requirements that address the needs and responsibilities of AI developers, AI deployers, and the Impact Population.

### 15.6.1 AI Developers

The role of the AI developers, as mapped across the paper, is predominant at the development stage, from ideation to deploying the AI solutions. AI developers' role is significant beyond the development stage as they directly/indirectly interface with the AI deployers who procure the AI solutions. Besides, one of the significant ways AI developers can contribute towards making Responsible AI is by tackling the potential impact that the technology could cause when deployed by the AI deployers or directly used by the Impact Population.

### 15.6.2 Plan & Design Stage

In this section, we will discuss various principles to be followed by the players, such as C-suite executives, Test & Evaluation, Validation & Verification experts, product managers, compliance experts, auditors, organisational management, etc. may follow to ideate AI solutions which are responsible and safe. In this stage, developers and technologists must focus on understanding their AI systems' potential consequences and implementing appropriate measures to mitigate risks through operationalising the following principles using the suggested strategies.

### 15.6.3 Collect and Process Data

During this stage, players such as Data scientists, data/model/system engineers etc., must carefully consider the principles and strategies to ensure responsible and ethical data practices by seeking diverse datasets representing different perspectives, demographics, and societal contexts. Adhering to these principles and employing the suggested strategies can enhance the reliability, fairness, and privacy of the data used in AI systems.

### 15.6.4 Build and Use Model

In this stage, AI developers (i.e., players like Modelers, Model Engineers, Data scientists, data/model/system engineers, domain experts, etc.) face the crucial task of carefully selecting suitable algorithms, building the model architecture, and establishing the specific techniques and methodologies to be employed. This stage is pivotal in achieving essential attributes such as robustness, explainability, fairness, generalisation, and privacy protection in the AI model's design. The thoughtful consideration of these factors ensures that the algorithm is effective, trustworthy, and aligned with responsible AI principles.

### 15.6.5 Verification and Validation

In the verification and validation stage in the AI lifecycle, developers and technologists (Data Scientists, experts etc.) delve deeper into ensuring the responsible and safe operation of AI systems before deployment. Building upon the principles outlined, this stage requires a meticulous focus on comprehending the potential consequences of AI systems and implementing effective risk mitigation measures. By overlaying the deployment context and making informed choices, developers can establish a robust foundation for successfully integrating AI systems while addressing potential risks and ethical concerns.

### 15.6.6 Deployment and Operationalisation

The deployment and operationalisation stage is crucial in operationalising AI principles. It entails deploying AI systems onto real products and their interaction with the environment and users. This stage focuses on fine-tuning the AI system to ensure its effectiveness and reliability in real-world scenarios. In this stage, AI Developers and technologists (Developers, System Engineers, Procurement experts etc.) work towards refining the system's performance, addressing any issues that arise, and optimising it for seamless integration into existing processes. The goal is to ensure that the AI system functions effectively and delivers the intended outcomes in real-world applications.

### 15.6.7 AI Deployers

AI deployers refer to individuals, organisations, or entities that utilise artificial intelligence solutions or systems in their operational processes. These users are the recipients or consumers of AI technology and leverage its capabilities to perform various tasks, make informed decisions, deliver services, or enhance their operations. AI deployers can span across different industries and sectors, such as healthcare, education, finance, manufacturing, law enforcement, and more. They interact with AI systems, either directly or indirectly, to leverage the outputs, insights, or recommendations generated by AI algorithms and models. AI deployers play a critical role in effectively implementing and utilising AI solutions, driving innovation, efficiency, and data-driven decision-making within their respective domains.

### 15.6.8 Impact Population

In the context of AI, the term "impact population" refers to the individuals or groups who are directly affected by the deployment and use of AI systems. The impact population includes the end-users, customers, or beneficiaries of AI applications, as well as any stakeholders who may be affected by the outcomes or consequences of the AI system. These individuals or groups may experience the direct impact of AI-generated decisions, services, or products.

## 15.7 Implementation of Principle-based Multistakeholder Approach

Coordination of various factors like regulatory landscape, geopolitics etc., is essential for the seamless implementation of the principle-based multistakeholder approach. In this section, we will discuss the government's role in implementing the principle-based multistakeholder approach by establishing different forms of coordination. While there are various levels at which India could need coordination to adopt a principle-based data multistakeholder approach, in this chapter, we will discuss three essential levels, i.e., Domestic Coordination, International Coordination, and Public-Private Coordination.

### 15.7.1 Domestic Regulatory Coordination

The zero step towards implementing the principle-based multistakeholder approach would require domestic stability in terms of regulations. The primary regulatory issue would be recognising this framework as a legitimate lens to establish responsible AI innovations in India. If the regulation and enforcement fall under the ambit of multiple regulators domestically, discussed in this section, recognition of this framework might not be uniform as some might recognise it while others refrain from it. In addition, the existence of different regulators/authorities will pave the way for multifarious interpretation/understanding of the framework, which gives birth to slightly different versions of the principle-based multi-stakeholder approach at the implementation level, causing confusion and conflict. Moreover, this conflict and differences at the implementation level will impact AI innovations, causing compliance uncertainty and regulatory arbitrage. Therefore, consistent recognition and implementation of a principle-based multi-stakeholder approach at domestic regulatory levels are crucial.

### 15.7.2 International Regulatory Cooperation

While domestic regulatory coordination is crucial, there are also various other roadblocks to implementing the principle-based multistakeholder approach towards the AI ecosystem, which can't be solved exclusively at the domestic level. A concerted effort is needed between India and other jurisdictions beyond its borders to make AI innovations responsible and safe. In an increasingly interconnected world, international regulatory cooperation has emerged as a crucial pillar of regulatory policy.[624] Various jurisdictions have also emphasised this in the context of AI governance, where they believe concerted international-level regulatory cooperation is the way forward.[625]

### 15.7.3 Establishing Public-Private Collaboration

Implementing the AI regulations is a fresh start for regulators and domestic industries in many jurisdictions, especially in the global south. The range of AI innovations to be tackled will be immensely vast, starting from big tech to Micro, Small and Medium Enterprises ('MSMEs') to government agencies. While a one-size-fits-all approach towards AI regulation might bring in compliance (at a cost) among the horizontally (AI general) and vertically (AI narrow) diverse range of AI developers and AI deployers, it might not bring cooperation. Therefore, governments must operationalise various market and regulatory mechanisms to build a healthy relationship and cooperation with AI developers and AI deployers with a limited disposal capacity.

The governments could follow normative theories of regulation[626] and institute market mechanisms such as a (a) audit of features for AI developers and AI deployers based on the principles mapped for them and (b) market for principles-based accreditation, enabling a competitive edge for platforms. While an independent auditing agency must perform the audit, a government or authorised entity must perform the accreditation process at a nominal cost based on defined principles. The accreditation process must have a well-laid process and procedure that balances transparency and safeguards to protect intellectual and proprietary information. Besides, the accreditation process must be aspirational such that it pushes the

[624] OECD. (2021). *Why does international regulatory cooperation matter and what is it?* OECD iLibrary. Retrieved June 20, 2023, from <https://www.oecd-ilibrary.org/sites/62c39d12-en/index.html?itemId=/content/component/62c39d12-en>.

[625] Kerry, C. F., Meltzer, J. P., Renda, A., Engler, A., & Fanni, R. (2022, March 9). *Strengthening international cooperation on AI*. Brookings. Retrieved June 20, 2023, from <https://www.brookings.edu/research/strengthening-international-cooperation-on-ai/>.

[626] UNESCO. (2021, November). *Recommendation on the ethics of artificial intelligence.* <https://en.unesco.org/about-us/legal-affairs/recommendation-ethics-artificial-intelligence>.

AI developers and AI deployers toward performing better on the user outcome aspect, i.e., securing the impact population from the adverse implications of AI technologies.

## 15.8 Conclusion

Humans are the heart of the Internet, and everyone should benefit from the open and trustworthy Internet. However, the Internet is going through a paradigm shift driven by key technological developments like Artificial Intelligence. These technological developments pose challenges to the internet at different levels, like (a) gaps in the regulatory parameters, (b) technological differences, (c) lack of interoperability for networking, (d) safety and security concerns impacting trust etc. These challenges directly implicate how humans perceive the Internet's future, which is currently filled with uncertainty, as highlighted by the previous version of the global Internet report.

Therefore, to transform the status quo, it is important to reinstate trust within disruptive technologies like Artificial Intelligence, which will fundamentally alter how we interact with the internet in the coming future. To achieve the same, there is a need for a governance framework which would enhance opportunities afforded by Artificial intelligence by making it trustworthy while minimising harm. Therefore, this is where our paper comes into the picture, adding value to efforts towards making AI development and deployment trustworthy by proposing an ecosystem-level principle-based approach which appropriately maps the harms and impact at the different stages and suggests principles for various stakeholders for tackling the same. Going further, this paper could set the context for future research on how the stakeholders can pragmatically put to action the identified principles and indicated operational strategies at scale.

# 16 Developing AI Standards that Serve the Majority World

**Michael Karanicolas, Executive Director, UCLA Institute for Technology, Law & Policy**

## Abstract

This essay considers the emerging transnational governance framework for AI that is being developed under the auspices of a handful of powerful regulatory blocs, which represent a relatively homogenous set of global interests. It will argue that, while attempts to develop binding rules of the road are laudable, the world would be better served if the standard-setting processes represented a more diverse set of stakeholders, and that perspectives from the people of the Majority World should be an essential component to developing new standards to govern the development and deployment of AI technologies.

## 16.1 Introduction

In 2017, a Nigerian Facebook employee named Chukweuemeka Afigbo shared a video of his struggles getting an automated soap dispenser to work, presumably because the machine's optic sensor was not calibrated to recognize his darker skin tone.[627] Ultimately, Mr. Afigbo ended up having to cover his hand with a white paper towel to get the machine to function. The post was captioned with a statement on the importance of diversity in the technology industry and the pitfalls of having a homogenous team working on products, especially where the makeup of this team may not reflect the realities of the product's usage in the real world.

While the consequences of failure in a case like this are relatively benign, a lack of representation can have serious repercussions. Over the past decade, there have been countless stories of products developed in Silicon Valley causing harm when introduced to

[627] Sidney Fussell, *Why Can't This Soap Dispenser Identify Dark Skin?,* Gizmodo (Aug. 17, 2017), <https://gizmodo.com/why-cant-this-soap-dispenser-identify-dark-skin-1797931773>.

diverse cultural, socioeconomic, or geographic contexts.[628] With the emergence of artificial intelligence (AI) as a major transformational technology, regulators around the world are determined to get ahead of the potential social harms by developing guardrails that are gradually coalescing into a new set of global standards for AI.[629]

There is no question that order and appropriately regulated growth is preferable to the chaos that dominated the previous era of disruption.[630] But just as context is important to the development of new technologies, regulatory standards which fail to reflect the localized impacts of a new technology can be ineffective, or even dangerous. A legal principle may have a certain meaning in, for example, a society where the rule of law and checks against administrative abuse are strong, and a completely different meaning in the context of a weak democracy or authoritarian government.[631] In any standards-development process, who gets a seat at the table is going to shape the values and priorities which underlie the final framework that emerges.

This essay considers the emerging transnational governance framework for AI that is being developed under the auspices of a handful of powerful regulatory blocs, which represent a relatively homogenous set of global interests. It will argue that, while attempts to develop binding rules of the road are laudable, the world would be better served if the standard-setting processes represented a more diverse set of stakeholders, and that perspectives from the people of the Majority World should be an essential component to developing

[628] *See, e.g.,* Paul Mozur, *A Genocide Incited on Facebook, With Posts from Myanmar's Military*, N.Y. Times (Oct. 15, 2018), <https://www.nytimes.com/2018/10/15/technology/myanmar-facebook-genocide.html, *archived* at <https://perma.cc/3Z2J-K6BA>; Kate O'Flaherty, *YouTube keeps deleting evidence of Syrian chemical weapon attacks*, Wired (Jun. 26, 2018), <https://www.wired.co.uk/article/chemical-weapons-in-syria-youtube-algorithm-delete-video>; Daniel Avery, *Gay Dating App Grindr Still Leaking Users' Location Data, Report Indicates*, Newsweek (Aug. 12, 2019), <https://www.newsweek.com/grindr-location-leak-1453697>.

[629] Daniel Schiff, Justin Biddle, Jason Borenstein, & Kelly Laas, What's Next for AI Ethics, Policy, and Governance? A Global Overview (AAAI/ACM Conference on AI, Ethics, and Society, 2020), <https://doi. org/10.1145/3375627.3375804>.

[630] *Supra* note 3.

[631] Jacob Mchangama & Natalie Alkiviadou, *The Digital Berlin Wall: How Germany (Accidentally) Created a Prototype for Global Online Censorship – Act Two*, Justitia (Sept. 2020),<https://justitia-int.org/wp-content/uploads/2020/09/Analyse_Cross-fertilizing-Online-Censorship-The-Global-Impact-of-Germanys-Network-Enforcement-Act-Part-two_Final-1.pdf>.

new standards to govern the development and deployment of AI technologies.[632]

The Essay proceeds in Part I by introducing AI and emerging conceptions of bias and other harms. Part II discusses the models of AI governance emerging globally, particularly from the United States, the European Union (EU), and China, which are best positioned to influence emerging global standards. Parts III and IV discuss the concept of regulatory diffusion and challenges to this model of international standard setting, before offering recommendations, in Part V, for a more inclusive model of development which takes into account the needs of diverse global stakeholders who are impacted by the transition to an AI society.

## 16.2 Understanding AI

AI refers to several technical concepts which can generally be understood under the umbrella of machine learning, which means that a system learns from data as opposed to following hard-coded rules. In essence, machine learning systems operate as statistical inference engines with the capacity to generate outputs from the analysis of large inputs of data.[633] However, the data-dependent nature of machine learning technology means that biases and errors can constantly leak into these processes, with the potential to automate and further entrench inequalities and inequities inherent in the social order from which the underlying data or development processes originate.[634] There can be a number of subtle avenues for

[632] *Terminology for how to distinguish between the world's high income and lower income economies is a fraught exercise, which is difficult to disentangle from the traditional colonial structure which undergirds terms like "the third world" or "the developing world". "Developing world" is particularly problematic, insofar as it paints a false picture of a narrowing gap between haves and have-nots, despite the fact that global inequities, and the exploitative relationships that reinforce these inequities, have proven extremely resilient. While "Global South" is a popular alternative term, it belies a perspective that is centered on the United States, Europe, and Canada. Australia and New Zealand, for example, are both paradoxically understood to be included within the Global North, while Mongolia and North Korea are Global South. All such binary distinctions are problematic insofar as they fail to grapple with the varying levels of development, income, and power around the world, as well as within countries at every development and income level. This essay will proceed to use the terms "Majority World" and "Minority World". Although this distinction inevitably glosses over important context, it is a useful reminder that the traditional geopolitical paradigm concentrates power and wealth in a minority of people at the expense of the majority. See generally* Shahidul Alam, *Majority World: Challenging the West's Rhetoric of Democracy*, 34 Amerasia Journal 88 (2008).

[633] Harry Surden, *Machine Learning and Law*, 89 Wash. L. Rev. 87, 87-90 (2014).

[634] Solon Barocas & Andrew D. Selbst, *Big Data's Disparate Impact*, 104 Cal. L. Rev. 671, 674 (2016).

these biases to take root, including as a result of the structure of the data fed into the system and the architecture of the algorithm.[635]

There is a voluminous literature on how problems, including biases, manifest, which offers potential responses aimed at countering these problems from a technical, social, and legal perspective.[636] Likewise, early scholarship has emphasized risks stemming from data sets that are either explicitly biased, or which are otherwise reflective of pervasive structural social problems.[637] Human biases can be introduced at every stage of the development and deployment process, even in unsupervised machine learning, based on how data is organized or success is defined.[638] All of these challenges are particularly severe in cases where there are significant geographic or cultural gaps between where models are built or where data is sourced and where the systems themselves operate.[639]

In addition to concerns about bias, accuracy, and efficacy, some leading scholars have asked more fundamental questions about AI's purported benefits and appropriateness. In "Automating Inequality," Virginia Eubanks poses two questions related to the basic ethics of AI deployment: (1) Does the tool increase the self-determination and agency of the poor? (2) Would the tool be tolerated if it was targeted at non-poor people?[640]

Across advanced democracies, however, the bulk of scholarship on this issue has focused on bias and discrimination, and problematic impacts

[635] *Ibid.* 716-722.

[636] *See, e.g.,* Ruha Benjamin, *Assessing Risk, Automating Racism: A health care algorithm reflects underlying racial bias in society,* 366 Science 421 (2019); Anupam Chander, *The Racist Algorithm?,* 115 Mich. L. Rev. 1023 (2017); Sonia K. Katyal, *Private Accountability in the Age of Artificial Intelligence*, 66 UCLA L. Rev. 54 (2019); Sandra G. Mayson, *Bias In, Bias Out*, 128 Yale L. J. 2218 (2019); Ifeoma Ajunwa, *The Paradox of Automation as Anti-Bias Intervention*, 41 Cardozo, L. Rev. 1671 (2020); Safiya Umoja Noble, Algorithms Of Oppression: How Search Engines Reinforce Racism (2018); Solon Barocas & Andrew D. Selbst, *Big Data's Disparate Impact*, 104 Cal. L. Rev. 671 (2016); Vidushi Marda & Shivangi Narayan, Data in New Delhi's Predictive Policing System (FAT '20: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, 2020) <https://doi.org/10.1145/3351095.3372865>.

[637] Kate Crawford & Ryan Calo, *There is a blind spot in AI research,* 538 NATURE, 311 (2016).

[638] James Zou & Londa Schiebinger, *Design AI so that it's fair,* 559 NATURE, 324 (2018).

[639] Shreya Shankar, Yoni Halpern, Eric Breck, James Atwood, Jimbo Wilson, & D. Sculley, *No Classification without Representation: Assessing Geodiversity Issues in Open Data Sets for the Developing World,* (NIPS 2017 Workshop on Machine Learning for the Developing Worldv, 2017).

[640] Virginia Eubanks, Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor (2018).

of these technologies on traditionally marginalized communities in the domestic context of the authors who are examining the issue.[641] As the next section demonstrates, this focus has colored regulatory priorities in the AI governance space.

## 16.3 Regulating AI

Around the world, the growing interest in AI has led to the emergence of number of different sets of principles, guidelines, or ethical codes that have been proposed or adopted.[642] However, relatively few governments have actually passed normative, hard law frameworks to govern this space.[643] Some of the most ambitious efforts aimed at developing a new regulatory structure for AI have emerged from the EU, particularly the proposed Artificial Intelligence Act (AI Act)., which focuses on potential risks of AI in terms of its security or potential to infringe on fundamental rights.[644]

The AI Act imposes a sliding set of requirements based on the purported risk of the application, such as obligations related to transparency, auditing, oversight, etc. Although the AI Act is the core of this new regulatory framework, other aspects of EU regulation, such as the General Data Protection Regulation and the Digital Services Act, are also relevant in setting standards for how AI systems must be developed and rolled out in certain contexts.[645]

In the United States the most high-profile attempt to impose uniform standards has been the Biden administration's Blueprint for an AI

[641] *This is not intended to overlook contributions from Majority World scholars to the current discourse, such as, for example,* Abeba Birhane, *Algorithmic Colonization of Africa*, 17 SCRIPTed 389 (2020); Vidushi Marda & Shivangi Narayan, Data in New Delhi's Predictive Policing System (FAT '20: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, 2020) <https://doi.org/10.1145/3351095.3372865>; Damian Okaibedi Eke, Kutoma Wakunuma, Simisola Akintoye, Responsible AI in Africa (2023); Arthur Gwagwa, Erika Kraemer-Mbula, Nagla Rizk, Isaac Rutenberg, Jeremy de Beer, *Artificial intelligence (AI) deployments in Africa: Benefits, challenges and policy dimensions*, 26 African Journal of Information and Communication 1 (2020).

[642] *See, e.g.,* https://au.int/sites/default/files/documents/38507-doc-dts-english.pdf; <https://unesdoc.unesco.org/ark:/48223/pf0000381137>.

[643] Blair Levin & Larry Downes, *Who Is Going to Regulate AI?,* Harvard Business Review (May 19, 2023), <https://hbr.org/2023/05/who-is-going-to-regulate-ai>.

[644] European Commission, 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 COM (2021) 206 final (Apr. 21, 2021) [hereinafter EU AI Act].

[645] *Id.*

Bill of Rights.[646] This framework shares the general thematic focus of the proposed AI Act, insofar as both are targeted at mitigating performance challenges, particularly decision-making which is biased, unfair, or otherwise erroneous. There are also significant structural differences, however, particularly as the Blueprint for an AI Bill of Rights is a non-binding set of guidelines. In January 2023, the National Institute of Standards and Technology released its own AI Risk Management Framework, a set of voluntary guidelines for organizations and individuals to support the trustworthiness of AI systems that they may be developing or deploying.[647] In addition, there have been a range of other executive orders aimed at tackling this challenge, along with legislation in several states, particularly California, Texas, Connecticut, and Illinois.[648]

The Organization for Economic Co-operation and Development (OECD), an intergovernmental organization comprised mainly of high-income economies, has also been a significant driver of international standards in this space, beginning with their Artificial Intelligence Principles, which were adopted in 2019.[649] The Global Partnership for Artificial Intelligence (GPAI), which was launched in 2020, further built on these principles, including through a set of "[p]rinciples for responsible stewardship of trustworthy AI" and "[n]ational policies and international cooperation for trustworthy AI".[650] The GPAI bills itself as a multistakeholder initiative, with avenues for participation by industry, civil society and independent experts. In practice, however, governments dominate the GPAI's structure and decision-making.[651] The GPAI Secretariat is hosted at the OECD, though it is open to non-OECD members, and at least four majority world countries have joined, namely Brazil, India, Senegal and Argentina.[652]

[646] *Blueprint for an AI Bill of Rights*, The White House, <https://www.whitehouse.gov/ostp/ai-bill-of-rights/>.

[647] *AI Risk Management Framework*, National Institute of Standards and Technology, <https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf>.

[648] *See, e.g.*, Exec. Order No. 13859, 84 FR 3967 (2019); Exec. Order No. 13960, 85 FR 78939 (2019); A.B. 331, 2023-2024 R. Sess. (Cal. 2023); ; S.B. 1103, Sess. Year 2023 (Conn. 2023); H.B. 3385, 103rd Gen. Assemb., 2023 and 2024, (Ill. 2023); H.B. 2060, 88th Leg., 2023-2024 (Tex. 2023).

[649] *OECD AI Principles Overview*, OECD Policy Observatory, <https://oecd.ai/en/ai-principles>.

[650] *About* GPAI, The Global Partnership on Global Intelligence, <https://gpai.ai/about/>.

[651] *Id.*

[652] <https://gpai.ai/community/>.

In recent years, China has been increasingly proactive in its attempts to establish itself as a hub for regulatory leadership and standard setting in AI. This includes efforts to empower national champions, especially Baidu, Alibaba, Tencent, Xiaomi (BATX), as well as concomitant efforts to drive standards through investment, particularly through the Digital Silk Road under the Belt and Road Initiative, which is framed as a South-South development alternative.[653] China, however, has also rolled out a number of groundbreaking policy initiatives, particularly through the powerful Cyberspace Administration of China, which recently imposed broad new rules to govern recommendation algorithms.[654] In contrast to rights-based or risk-based approaches to AI, China's regulatory landscape draws heavily from cybersecurity structures, which amalgamate conceptions of data security with a broader focus on national security.[655]

Thematically, although questions like enforceability vary across jurisdictions, governance efforts in the United States, the EU, and the OECD tend to revolve around solutions aimed at combatting the perceived harms of AI applications, including product liability rules, data privacy rules, safety standards, requirements related to explainability and fairness, and, in some instances, outright prohibitions on the uses of AI systems for particularly problematic purposes.[656] China's rules, while somewhat more focused on security and order, also cover many of these same areas, notably related to explainability, trustworthiness, oversight, and broader ethical norms for developing and using AI.[657]

[653] Matthew S. Erie & Thomas Streinz, *The Beijing Effect: China's 'Digital Silk Road' as Transnational Data Governance*, 54 N.Y.U. J. Int'l L. & Pol. 1 (2021); Marie-Therese Png, At the Tensions of South and North: Critical Roles of Global South Stakeholders in AI Governance (ACM Concerence on Fairness, Accountability, and Transparency, 2022), <https://dl.acm.org/doi/10.1145/3531146.3533200>.

[654] Micah Musser, *Don't Assume China's AI Regulations Are Just a Power Play*, Lawfare (Oct. 3, 2022), <https://www.lawfareblog.com/dont-assume-chinas-ai-regulations-are-just-power-play>.

[655] *Emotional Entanglement: China's emotion recognition market and its implications for human rights*, Article 19 (Jan. 2021), 46-48, <https://www.article19.org/wp-content/uploads/2021/01/ER-Tech-China-Report.pdf>.

[656] Emma Carmel & Regine Paul, *Peace and prosperity for the digital age? The colonial political economy of European AI governance*, 41 IEEE Technol. Soc. Mag. 94, 96 (2022).

[657] Matt Sheehan, *China's New AI Governance Initiatives Shouldn't Be Ignored*, Carnegie Endowment for International Peace (Jan. 4, 2022), <https://carnegieendowment.org/2022/01/04/china-s-new-ai-governance-initiatives-shouldn-t-be-ignored-pub-86127>.

The examples mentioned above are not the only AI regulatory efforts. Other noteworthy examples include Canada's Directive on Automated Decision-Making, which governs the development and deployment of AI systems across that country's federal agencies.[658] Singapore has also been an early mover in this space, through the launch of its own Model AI Governance Framework and, more recently, the development of A.I. Verify, a testing framework toolkit designed to support independent self-assessment by private sector actors developing or employing AI technologies.[659]

These initiatives, however, are clustered in the minority world.[660] Moreover, while frameworks such as Canada's and Singapore's represent important contributions to the global discourse on regulating AI, they lack the institutional support to drive broader standard setting in the way initiatives based in the United States, the EU, and China are able to. The next Part discusses standard setting as a general phenomenon and introduces the drivers and origins for how a framework gains international legitimacy.

## 16.4 Regulatory Diffusion

While there are substantial differences between the governance standards being considered across the advanced economies mentioned in the previous section, there is also significant overlap, representing a consolidation around particular understandings of the challenges inherent in AI and the appropriate scope of regulatory responses. Collectively, this emerging consensus, which has been driven largely by frameworks developed among wealthy and powerful countries, leads to external pressure on other countries to either adopt a similar regulatory framework, or to cede their regulatory position on these issues altogether, a phenomenon which is sometimes referred to as "regulatory diffusion".[661] While the "Brussels Effect" is

[658] http://www.tbs-sct.gc.ca/pol/doc-eng.aspx?id=32592&section=html.

[659] *Singapore's Approach to AI Governance*, Personal Data Protection Commission (May 2022), <https://www.pdpc.gov.sg/Help-and-Resources/2020/01/Model-AI-Governance-Framework>.

[660] *See, e.g.,* Chukwubuikem I. Obianyo & Solomon Vendaga Ater*, A Critical Appraisal of the Legal Framework of Artificial Intelligence Governance in Nigeria, 4* Journal Of Private And Public Law 48 (2023), *discussing the regulatory gap in Nigeria.*

[661] Jennifer Nou and Julian Nyarko, Regulatory Diffusion, 74 Stanford Law Review 897 (2023).

probably the best-known framing for how local standards become globally influential, these impacts are not limited to EU processes.[662]

There can be a number of drivers which inspire countries to copy or adopt laws or legal principles from elsewhere, including the efficiency of harmonized regulations, or even simply to save the resources required to develop their own approach.[663] This trend can be particularly powerful in the context of emerging democracies, which often turn to more established democracies to build legitimacy behind a particular course of action.[664] Similar tendencies, however, can play out across more authoritarian models of governance, as evidenced by the rash of new criminal misinformation laws and misinformation prosecutions of journalists and opposition figures that accompanied the COVID-19 pandemic.[665]

Regulatory diffusion can be a positive phenomenon, such as the rapid global proliferation of freedom of information (or right to information) legislation which has taken place since the 1990s.[666] Although there were certainly coercive elements at play in this process, such as the use of foreign aid or multilateral institutions to pressure countries into adopting these laws as a mechanism for democratic accountability and as a check against corruption, the end result has been broadly beneficial from the perspective of human rights and democracy.[667] Similar diffusion pressures have been observed related to a number of other constitutional rights.[668]

Regulatory diffusion is not a universally positive phenomenon. Political scientists have noted that, while adoption based on learning about effective policies elsewhere can provide for good outcomes,

[662] Annegret Bendiek & Isabella Stuerzer, *The Brussels Effect, European Regulatory Power and Political Capital: Evidence for Mutually Reinforcing Internal and External Dimensions of the Brussels Effect from the European Digital Policy Debate*, 2 Digital Society 5 (2023).

[663] Jonathan M. Miller, *A Typology of Legal Transplants: Using Sociology, Legal History and Argentine Examples to Explain the Transplant Process*, 51 Am. J. Comp. Law 839, 846 (2003).

[664] *Ib.*

[665] Michael Karanicolas, *Even in a Pandemic, Sunlight Is the Best Disinfectant: COVID-19 and Global Freedom of Expression*, 22 Oregon Review of International Law 101 (2021).

[666] Margaret Kwoka & Michael Karanicolas, *Overseeing Oversight*, 54 Conn. L. Rev. 657, 663 (2022).

[667] *See, e.g., General Comment No. 34: Article 19 (Freedoms of opinion and expression)*, UNHRC, 102nd Sess, UN Doc CCPR/C/GC/34 (2011); *Claude Reyes and Others v Chile* (2006) Inter-Am Ct HR, (Ser C) No 151.

[668] Benedikt Goderis & Mila Versteeg, *The Diffusion of Constitutional Rights*, 39 Int'l Rev. L. Econ. 1 (2014).

diffusion can also occur based on competition. This can occur where a government faces economic pressure to ensure that their regulatory framework is as attractive to prospective investors as their peers, or even through direct coercion by more powerful governments.[669] Both of these mechanisms are likely to produce regulatory postures which fail to optimally serve the needs of locals.[670] For example, concerns over such pressures among the states was a factor underlying the adoption of the commerce clause in the U.S. Constitution.[671]

## 16.5 AI Governance as a Standard Setting Exercise

In its 2022 *AU Data Policy Framework*, the African Union urged Member States to adopt a coordinated, comprehensive and harmonized regional approach to global digital governance challenges, including with regards to technical standards, ethics, governance, and best practices related to AI.[672] Governments in the majority world, however, face significant obstacles to developing independent AI governance frameworks which suit the needs of their constituents. First is the simple challenge of compelling compliance. Companies that are on the leading edge of AI development tend to be headquartered in high income countries, leaving poor countries with far less leverage in influencing the companies' decision-making.[673] Outside of a handful of particularly large markets, such as Brazil or India, Majority World countries face a binary choice between accepting the inherent problems or biases in these technologies or foregoing their associated economic benefits entirely and risk being left behind.

Relatedly, and as noted in the previous section, significant intergovernmental momentum has already built behind the frameworks that have been developed by advanced economies. This leads to direct pressure on majority world governments to join existing initiatives, such as the one being pushed by the OECD. For example, the "Egyptian Charter on Responsible AI", which was published in

[669] Charles R. Shipan & Craig Volden, *The Mechanisms of Policy Diffusion*, 52 AM. J. POL. SCI. 840, 841-848 (2008).

[670] *Id.*

[671] *Id.* at 849.

[672] <https://au.int/sites/default/files/documents/42078-doc-AU-DATA-POLICY-FRAMEWORK-ENG1.pdf,%20p.%2043-44>.

[673] Abeba Birhane, *Algorithmic Colonization of Africa*, 17 SCRIPTed 389, 392-393 (2020).

2021, draws heavily from the OECD principles.[674] Countries must accept the framing and perspective that underlies these projects in order to have a seat at the table going forward.

In this context, it should not be surprising that the major global governance frameworks which have emerged are generally focused on impacts across a set of prioritized stakeholders. For example, where the EU's proposed AI Act contemplates which uses of AI should be fully prohibited, the focus is on subliminal manipulation, exploitation of vulnerable people, general purpose "social credit scoring", and real-time biometric identification.[675] The latter prohibition, however, is subject to limited exceptions based on public safety threats. This carveout may be suitable in the context of a country like France or Germany, which is relatively stable and has robust protections for democracy and the rule of law.[676]

In the context of a country like Uganda or Nigeria, such a loophole is likely to be abused, due to the authoritarian tendencies of their leadership, the lack of strong protections for broader democratic rights, and a more precarious security situation across the country.[677] Underlying ethnic or political tensions, and the likelihood of mass violence, should also impact the calculus for whether an AI-driven tracking or surveillance program may be acceptable if subjected to careful safeguards, or whether it should be prohibited entirely.

Frameworks which originate in wealthier states often fail to fully grapple with concerns that global AI supply chains will throttle the potential for homegrown technological development in poor regions, presenting an obstacle to equitable development as increasing shares of the economy are transformed by AI.[678] The exploitative labor relationships which underlie the development and improvement of

[674] <https://aicm.ai.gov.eg/en/Resources/EgyptianCharterForResponsibleAIEnglish-v1.0.pdf>.

[675] EU AI Act, *supra* note 14, title II, article 5.

[676] *See, e.g., Freedom House's annual index of democratic health, which rank nearly all EU countries at or near the top of the list: https://freedomhouse.org/countries/freedom-world/scores?sort=desc&order=Total%20Score%20and%20Status.*

[677] *See Uganda: Events of 2021*, Human Rights Watch, <https://www.hrw.org/world-report/2022/country-chapters/uganda>; *Nigeria: Events of 2021*, Human Rights Watch, <https://www.hrw.org/world-report/2022/country-chapters/nigeria>.

[678] Arthur Gwagwa, Erika Kraemer-Mbula, Nagla Rizk, Isaac Rutenberg, Jeremy de Beer, *Artificial intelligence (AI) deployments in Africa: Benefits, challenges and policy dimensions*, 26 African Journal of Information and Communication 1, 4 (2020); Abeba Birhane, *Algorithmic Colonization of Africa*, 17 SCRIPTed 389, 396 (2020).

AI systems, or the toxic and harmful impacts of extractive industries which are designed to provide energy or raw materials for their production, are also generally not areas of priority.[679]

AI research and development is enormously energy-intensive, compounding and accelerating climate change threats which will be disproportionately borne by residents of the majority world.[680] AI development is also undergirded by extractive supply chains whose environmental impacts are likewise centered in poor countries.[681] The development of AI requires enormous amounts of labor to label datasets, curate and moderate harmful content, and train and input data, which is likewise typically drawn from the global poor.[682]

While it may seem intuitive to many public policy professionals in the minority world to separate discussions about AI fairness and privacy from environmental or labor concerns related to the development of these sectors, it is likely not coincidental that this division lines up with a geographic delineation in how the harms from AI manifest. It is also worth noting that the piloting of AI technologies across the EU and North America often targets disempowered populations, including data subjects from the majority world, such as through the prevalence of AI technologies in the EU's migration system.[683]

## 16.6 Governing for the Majority

There are various existing avenues for governance conversations which allow for representatives from the majority world to address these issues on a more equal footing with their more economically advanced counterparts. The International Telecommunications Union (ITU), for example, provides a platform for discussions related to

[679] Carmel, *supra* note 26.

[680] Mark Coeckelbergh, *AI for climate: freedom, justice, and other ethical and political challenges* 1 AI and Ethics 67 (2021); Payal Dhar, *The carbon impact of artificial intelligence*, 2 Nature Mach. Intell. 423 (2020).

[681] Danae Tapia & Paz Peña, *White gold, digital destruction: Research and awareness on the human rights implications of the extraction of lithium perpetrated by the tech industry in Latin American ecosystems*, Global Information Society Watch (2020), <https://giswatch.org/node/6247>.

[682] Kate Crawford, Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence (2021), 53-89.

[683] Julien Jeandesboz, Technology, knowledge and the governing of migration, *in* Handbook on the Governance and Politics of Migration 337 (E. Carmel, K. Lenner, and R. Paul, Eds. 2021).

inclusive development of AI technologies and equitable access to their benefits.[684]

As a United Nations specialized agency with 193 member states, this structure is naturally more inclusive than the OECD, or purely domestically driven frameworks that do not account for the majority world at all.[685] However, it still fails to address challenges of inclusion and opacity, since this dynamic may not capture the nuances of the relationship between governors and governed, and exploiters and exploited. As Chinmayi Arun points out in her chapter for the Oxford Handbook of Ethics of AI, on *AI and the Global South: Designing for Other Worlds*, a temptation to view these challenges as part of a binary relationship between the developed and developing world is problematically reductionist.[686] While traditional colonial extractive and exploitative relationships certainly exist, the story of AI's diffusion across the majority world also includes cases such as India's Aadhar biometric database, which was driven by a political and industrial elite within that country to force the marginalized into a pervasive system of surveillance, as well as to systematically deny them other rights.[687] The emergence of China as a hub for the sale of abusive surveillance technologies to countries like Ethiopia, Brazil, Ecuador, and Kenya further complicates the narrative.[688] This is not to gloss over the traditional and ongoing role of European and U.S.-based companies in the global spyware trade.[689] However, as Arun notes, "institutional frameworks of Southern countries must be taken into account as we consider what impact AI might have on the South… The rights of Southern populations can be realized through efforts made by states, but can also be eroded by the governing elite of states."[690]

It is certainly true for a broad cohort of countries that their relationship with AI is dominated by their role in the supply chain: providing raw

[684] *Artificial* Intelligence, International Telecommunication Union, <https://www.itu.int/en/action/ai/Pages/default.aspx>.

[685] *See* https://www.itu.int/hub/membership/how-we-are-governed/.

[686] Chinmayi Arun, *AI and the Global South: Designing for Other Worlds*, *in* The Oxford Handbook of Ethics of AI (Markus D. Dubber, Frank Pasquale, & Sunit Das, eds. 2019).

[687] *Id. at 7-8.*

[688] *Id. at 9.*

[689] <https://www.privacyinternational.org/sites/default/files/2017-12/global_surveillance_0.pdf, p. 18>.

[690] *Id. at 12.*

materials for export, as well as data for companies based in wealthier parts of the world to extract in order to improve their products.[691] This dynamic, however, typically takes place with the acquiescence of local governments who may contract with the companies to provide public services, or otherwise demand access to the data collected as part of the cost of carriage.[692]

From a governance perspective, challenges in ensuring robust representation through governments necessitates that emerging AI standards be considered along a multidimensional axis. Beyond a myopic focus on risks to data subjects, or even a geopolitical context of rich countries and poor countries, the development and deployment of these technologies must be subject to a holistic assessment of impacts across a range of different stakeholder groups. As problematic as it is for a small cadre of decision-makers in Washington D.C., San Francisco, or Brussels to develop standard setting processes that will guide global AI development, extending these processes to include small numbers of elite representatives from industry or governments in the majority world is only a marginal improvement.[693]

Instead, the development of standards that reflect the needs of these diverse stakeholders requires an approach which goes beyond traditional governmental policymaking. Recent years have seen a number of experiments in new forms of governance, particularly clustered in the tech space. These have included the Global Internet Forum to Counter Terrorism (GIFCT), an industry-led self-regulatory initiative which works to set content standards for participating social media platforms, including through the development of machine-learning algorithms to catch extremist content and a shared hash

[691] Paola Ricaurte, *Data Epistemologies, Coloniality of Power, and Resistance*, 20 Television & New Media 350, 358 (2019).

[692] *Id.*

[693] For a more recent example, see Ian Bremmer and Mustafa Suleyman's September 2023 article in Foreign Affairs, which essentially proposes folding major technology companies into the global governance space. While Bremmer and Suleyman are correct on the need for industry buy-in and technical expertise to support regulatory conversations in this space, the involvement of major tech players is not itself sufficient to guarantee that emerging frameworks reflect the interests of those on the sharpest edge of technological change. Ian Bremmer and Mustafa Suleyman, The AI Power Paradox Can States Learn to Govern Artificial Intelligence—Before It's Too Late?, 102:5 Foreign Affairs (2023).

database.[694] GIFCT was designed to foster collaboration between governments, the private sector, and civil society, though the latter has complained of a lack of transparency.[695] Facebook's moves to empower an Oversight Board to review content decisions is also worth noting, insofar as it represents a (limited) derogation of power from the corporation to an arm's length entity.[696] Though the Oversight Board is not technically a multistakeholder body, it has included significant engagement with civil society.[697]

Probably the most well-established example of actual multistakeholder governance is the Internet Corporation for Assigned Names and Numbers (ICANN), a non-profit corporation that oversees a number of critical technical functions underlying the global internet, including managing the generic top-level domain name system ("gTLD") and the country code top-level domain name system ("ccTLD").[698] ICANN's decision-making takes place across multiple layers, led by a president and a board of directors, along with a number of other diffuse decision-making bodies which focus on particular areas or subthemes.[699] ICANN's multi-stakeholder model includes spaces for engagement by governments through the Governmental Advisory Committee,[700] engagement by civil society through the Non-Commercial Stakeholder Group,[701] engagement by internet end users through the At-Large Advisory Committee,[702] and engagement by business interests through the Commercial Stakeholders Group.[703]

[694] Hash Sharing Consortium, Global Internet Forum to Counter Terrorism, <https://gifct.org/joint-tech-innovation>.

[695] Emma Llansó, *Platforms Want Centralized Censorship. That Should Scare You*, Wired (Apr. 18, 2019), <https://www.wired.com/story/platforms-centralized-censorship/>.

[696] *Oversight Board Bylaws*, Facebook (Jan. 2020), <https://about.fb.com/wp-content/uploads/2020/01/Bylawsv6.pdf>.

[697] Brent Harris, *Preparing the Way Forward for Facebook's Oversight Board*, Facebook (Jan. 28, 2020), <https://about.fb.com/news/2020/01/facebooks-oversight-board.

[698] *Welcome to ICANN!*, ICANN (Nov. 8, 2019), <https://www.icann.org/resources/pages/welcome-2012-02-25-en>, *archived at* <https://perma.cc/M8T9-2XCS>.

[699] *See Insperity OrgPlus 2012*, ICANN (Nov. 3, 2019), <https://www.icann.org/en/system/files/files/management-org-01may18-en.pdf>, *archived at* <https://perma.cc/4TRW-Z83N>.

[700] *See Governmental Advisory Committee*, ICANN Governmental Advisory Committee (Nov. 8, 2019), <https://gac.icann.org/>, *archived at* <https://perma.cc/Z4LT-3MCY>.

[701] *See Non-Commercial Stakeholder Group*, ICANN (Nov. 8, 2019), <https://gnso.icann.org/en/about/stakeholders-constituencies/ncsg>, *archived at* <https://perma.cc/KW4Y-45X5>.

[702] *See About Us*, ICANN At-Large, <https://atlarge.icann.org/about/index, *archived at* <https://perma.cc/3L22-XZ5P>.

[703] *See Commercial Stakeholder Group*, ICANN, <https://gnso.icann.org/en/about/stakeholders-constituencies/csg>, *archived at* <https://perma.cc/CT4R-FM77>.

There is also a heavy emphasis on engagement and representation across regions.

Structurally, an independent multistakeholder AI governance body could act as a central hub for convening and policymaking by expert thematic subgroups, supported by robust public consultation and engagement processes. It could also support research, particularly by allowing secure sharing of information across companies and between companies and accredited researchers. It is worth noting that such a framework for information sharing is currently contemplated by the EU's Digital Services Act, although the ambition of this plan is limited by its thematic and geographic focus.

While ICANN's ability to retain its legitimacy as a hub for policy development in the domain name space shows that multistakeholder collaboration is possible in a manner which is not unduly dominated by nation-states, the organization has faced its share of criticisms and challenges. In addition to broader concerns about accessibility, there have been criticisms that the structure is not as egalitarianism as it claims, with particular risks of capture by commercial players, whose resources allow them to find ways to tilt the playing field in their favor even in the context of a consensus-driven and multistakeholder process.[704] It is also worth noting that ICANN's legitimacy emerged from a relatively unique set of circumstances, for which there is no parallel in the AI governance space.[705] ICANN's remit is also narrow and relatively technocratic, compared to the thematically sprawling and politically controversial world of AI governance.

There is also a tension between harmonized standards and the "hyper-local" way in which algorithmic harms manifest, which suggests a need for localized responses to mitigate these harms.[706] Any set of global, or even regional, standards, is bound to gloss over important contextual cues related to the specific cultural, linguistic, political,

[704] *See, generally,* Michael Karanicolas, *The New Cybersquatters: The Evolution of Trademark Enforcement in the Domain Name Space,* 30 Fordham Intellectual Property, Media & Entertainment Law Journal 399 (2020) (discussing how IP interests have had an outsized impact on the development of trademark policy in the domain name space).

[705] *See* Milton L. Mueller, *Detaching Internet Governance from the State: Globalizing the IANA*, 4 Geo. J. Int'l Aff. 35 (2014).

[706] Chinmayi Arun, *Rebalancing Regulation of Speech: Hyper-Local Content on Global Web-Based Platforms*, Medium (Mar. 28, 2018), <https://medium.com/berkman-klein-center/rebalancing-regulation-of-speech-hyper-local-content-on-global-web-based-platforms-1-386d65d86e32>.

or social nature of AI's impact in a given place or time.[707] A natural objection to calls for new multistakeholder body to develop AI governance standards is to query whether a centralized approach is desirable at all, or whether the inefficiencies of a patchwork of local rules are a worthwhile price to pay if it ensures that the rules appropriately reflect each unique local context.

Either way, the world faces a pressing need to ensure that the interests of stakeholders who are on the frontlines of AI's global impact are reflected in how these technologies are governed. Standard setting, and clear and binding policy, are desirable outcomes. The concern that AI is replicating traditional biases, inequities and discrimination within the societies where it has been developed, is well-grounded. However, it is critical that new governance structures aiming to mitigate these challenges do not themselves reflect traditional colonial contexts that have been the source of so much of the world's poverty, oppression and inequity.[708]

## 16.7 Conclusion

For seventy years, researchers studying automobile safety primarily based their work on the use of crash test dummies that were designed around what the industry considered to be the default dimensions of European and American men.[709] Because this research drove the development of vehicles' safety features, it led to design choices which supported favorable crash survival outcomes among this demographic, at the cost of worse survival rates among those with different body types.[710] Context matters, and a lack of appropriate context can have dire, even fatal consequences for those unlucky enough to be excluded from consideration.

[707] Arthur Gwagwa, Erika Kraemer-Mbula, Nagla Rizk, Isaac Rutenberg, Jeremy de Beer, *Artificial intelligence (AI) deployments in Africa: Benefits, challenges and policy dimensions*, 26 African Journal of Information and Communication 1, 8-10 (2020).

[708] Abeba Birhane, *Algorithmic Colonization of Africa*, 17 SCRIPTed 389, 391-393 (2020);.

[709] Tao Xu, Xiaoming Sheng, Tianyi Zhang, Huan Liu, Xiao Liang, & Ao Ding, *Development and Validation of Dummies and Human Models Used in Crash Test*, Applied Bionics and Biomechanics (2018), <https://www.hindawi.com/journals/abb/2018/3832850/.

[710] *Injury Vulnerability and Effectiveness of Occupant Protection Technologies for Older Occupants and Women*, National Highway Traffic Safety Administration (May 2013), <https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/811766>.

It is a good thing that the world's most influential policy-makers appear to be taking a proactive approach towards AI regulation, and there is no question that harmonization has its advantages, particularly in a globalized world. Governance standards which seek to bolster the rights of those who are adversely impacted by AI in the context of advanced economies are laudable. But as these frameworks begin to coalesce into transnational standards, it is important to query whether they actually represent the needs and concerns of those on the sharpest edge of technological disruption, or whether such global standards are seeking to address traditional domestic inequities by further entrenching inequities on a global scale.

Policy-makers, particularly across the world's advanced economies, should view the current moment as an opportunity to develop a stronger model of multistakeholder governance that establishes robust normative and ethical guardrails against harmful impacts of AI, particularly as experienced in the Majority World. Such collaborations are never easy, and asking politicians from advanced economies to expand their prioritization beyond the interests of their own constituents is particularly challenging. However, the disastrous consequences of the past two decades of technologically enabled disruption provide ample proof of the need for an inclusive approach to addressing the next generation of harms.

# PART 5

## Latin American Perspectives on AI Governance

# 17 AI Development Model for the Brazilian Justice Ecosystem: A Case study on the Operational Artificial Intelligence Sandbox Experience at the Public Defender's Office of Rio de Janeiro (DPRJ)

**Pedro Braga, Institute for Technology & Society (ITS Rio); Christian Perrone, Institute for Technology & Society (ITS Rio).**

## Abstract

This paper delves into the formation of a secure and ethical artificial intelligence (AI) framework for the Brazilian public sector in order to propose guidelines to structure innovative models of AI development. Collaborative efforts involving the Public Defender's Office of Rio de Janeiro, civil society groups, academics, development technology companies, and the Institute for Technology & Society of Rio de Janeiro (ITS Rio) have laid the groundwork testing an inclusive AI development model, addressing the needs of marginalised communities and vulnerable groups. Leveraging Brazil's extensive legal data landscape and the pioneering spirit of public defenders in adopting digital tools, the study employed machine learning techniques to amplify the analysis of judicial data and propose mechanisms to develop an AI for public policy. Through enhanced data analysis, even simple AI solutions can offer profound insights and operational efficiency. The paper highlights the successful implementation of an Operational AI Sandbox approach, ensuring the responsible development of technology in the public sector. It showcases how challenges in terms of participation, representation and ethical risk mitigation were addressed and expands on how the same model can be applied in other situations. Specifically, the development model used a Multistakeholder Committee strategy encompassing diverse perspectives, to establish ethical guidelines and principles for AI tool development and to monitor and test its actual implementation. This article shares insights gained from the collaborative process, presenting a potential AI model for the public sector. By examining the DPRJ experience, the study shares its step-by-step approach and key takeaways.

## 17.1 Introduction

How to build secure and responsible Artificial Intelligence tools (AIs) in the Public Sector? The challenges seem to be multiple, particularly in terms of participation, representation, and inclusion. A solution indeed seems to demand a review of the current processes that tend to focus on involving solely the client (public body in need of an AI tool) and the developers (AI based technology companies). The risks seem to be high particularly for vulnerable individuals and marginalised communities. The proposed path towards an AI development model may come from two different areas: innovative development methodologies (in this case AI operational sandboxes) and multistakeholder governance approaches (a multistakeholder Committee).

In a partnership with the Public Defender's Office of Rio de Janeiro (DPRJ), a novel AI public sector development model was proposed.[711] Using open judicial data, the pilot project aimed to enhance the work of the DPRJ (Public Defender's Office of Rio de Janeiro) in the realm of access to healthcare. In light of the remarkable results of the last 20 years that Public defenders have helped to secure the right to health,[712] an increased social demand has created strains in the human resources of the institution. Currently, there is one public defender for every 150,000 people[713] and the number of cases involving denied access to medicine has increased by about 5% each year, with at least 500,000 cases still pending.[714] Additionally, cases related to denied access to medicine in the city of Rio de Janeiro alone surpass 100 cases per month, with peaks of 10,000 cases per month in the whole state.[715]

[711] The Future of AI in The Brazilian Judicial System. (n.d.). ITS Rio. Retrieved August 24, 2023, from <https://itsrio.org/en/publicacoes/the-future-of-ai-in-the-brazilian-judicial-system/>.

[712] DPRJ. (2022, July 21). Histórias do Plantão Noturno: defesa do direito à saúde é destaque. Www.defensoria.rj.def.br. <https://www.defensoria.rj.def.br/noticia/detalhes/20377-Historias-do-Plantao-Noturno-defesa-do-direito-a-saude-e-destaque>.

[713] DPRJ. (2020, December). Relatório sobre o perfil dos réus atendidos nas audiências de custódia no período de agosto a dezembro de 2020 (pp. 1–39) [Review of Relatório sobre o perfil dos réus atendidos nas audiências de custódia no período de agosto a dezembro de 2020]. Diretoria Pública do Estado do Rio de Janeiro. Retrieved August 24, 2023, from <https://www.defensoria.rj.def.br/uploads/arquivos/09d3bcf2aa2c44e28fb55498d0a65f3d.pdf>.

[714] Idem.

[715] Idem.

One should note that citizens represented by public defenders tend to be the most vulnerable, especially living in the least affluent areas, a high portion in the slums (*favelas*) of the state of Rio de Janeiro. For instance, 81% of the individuals assisted by the DPRJ have a household income of up to one minimum wage.[716]

The use of data to enhance the judicial system is a well-known case in Brazil, a country that possesses the largest digital collection of open judicial data in the world.[717] The DPRJ, in this regard, is among the pioneers in developing innovation teams to use digital tools more efficiently. Utilising machine learning techniques, therefore, can drastically improve the analysis of judicial data, providing unique insights and increasing work efficiency, even when the applied AI is simple and accessible.

It was based on this insight that the project sought to delve into litigation data in order to create a tool that could support the work of public defenders. The idea was to streamline the defenders work, clustering similar cases and facilitate public policy negotiations with the defendants. The testing ground was access to health, more specifically access to medicine. Two main issues were raised: how to ensure the ethical development of the AI tool and the participation and representation of the impacted population.

The strategy was to use an Operational AI Sandbox methodology, whose primary purpose is to test novel technologies under a controlled and secure environment, and to set up a Multistakeholder Committee encompassing different perspectives to construct an inclusive artificial intelligence tool, guided by ethical principles and guidelines.

This paper discusses the outcomes of this process, culminating in a potential AI development model for the Brazilian public sector, rooted in this experience of an Operational AI Sandbox for the DPRJ. Lessons learned in constructing ethical and responsible AI are shared alongside a step-by-step approach, obtained through a case study in partnership with the institution's team. The data obtained

[716] Idem.

[717] This data publicly available and updated on a yearly basis by the Brazilian National Council of Justice (CNJ), on the open platform *Justiça in Números* (Justice in Numbers), available at: <https://justica-em-numeros.cnj.jus.br/> (Access on September 12, 2023).

can additionally provide insights towards better access to health in the state of Rio de Janeiro.

## 17.2 Justification

Brazil boasts the world's largest judicial system,[718] both due to its core activity, which involves judicially safeguarding individual, collective, and social rights and due to its high rates of open data.[719] According to the Judges' Productivity Index,[720] in 2021 alone, around 6 cases of judicialization were adjudicated per business day, totalling over 1,580 cases and amounting to 26.9 million judgments across Brazil. Most of these cases are available as open data resources, although not necessarily in a structured manner.

Since 2018, the use of artificial intelligence has gained momentum within this ecosystem due to the challenges posed by digitization and the dynamic nature of electronic processes within the Brazilian Judiciary. This trend is highlighted in the report titled "Exchange of Experiences between the European Union and Brazil on E-Justice."[721] In the years following, the Brazilian Judiciary has consistently expanded its investments in artificial intelligence. According to a recent report by Fundação Getúlio Vargas, half of the country's courts have already embraced this technology.[722] As an example, the development of the national platform for the management and training of AI models, Sinapses[723] stands out. Because it is a platform that is both managed by the CNJ from a central hub in Brasilia but also open to participation in governance by the various state courts, Sinapses supports the strategy of continuous sharing and innovation,

[718] The Future of AI in The Brazilian Judicial System. (op. cit.).

[719] Andrade, P. (2022, September 5). Justiça em números 2022: cada magistrado julgou 6,3 processos por dia útil em 2021. AMB. <https://www.amb.com.br/justica-em-numeros-2022-cada-magistrado-julgou-63-processos-por-dia-util-em-2021/>.

[720] Idem.

[721] CNJ. (2022). "Intercâmbio de experiências entre a União Europeia e o Brasil sobre e-Justiça [Review of "Intercâmbio de experiências entre a União Europeia e o Brasil sobre e-Justiça]. Conselho Nacional de Justiça. <https://www.cnj.jus.br/wp-content/uploads/2022/09/seminario-e-justice-v6.pdf>.

[722] Salomão, L. F. (2022). Artificial intelligence: technology applied to conflict management within the Brazilian Judiciary. Bibliotecadigital.fgv.br. <https://bibliotecadigital.fgv.br/dspace/handle/10438/33954?locale-attribute=en>.

[723] <https://www.cnj.jus.br/sistemas/plataforma-sinapses/> (Access on September 12, 2023).

while also preventing technological disparities among the courts. Simultaneously, it fosters integration within the Judiciary.

Alongside the Courts, other actors within the justice ecosystem have intensified their efforts in developing AI solutions, including Public Defenders. There is a goal to design ethical and responsible AI development models that promote human rights and foster participation, minimising risks and potential discrimination and exclusion.

The account of the experience in constructing this ethical and responsible AI development model, utilising the Operational Sandbox methodology, brings together not only the steps followed by ITS researchers and the technical team and employees of DPRJ but also the best practices gained through comparative studies and literature reviews. These practices can serve as inputs and references for crafting public policies in the field.

## 17.3 AI Sandboxes as a Methodology for Technological Development

Sandboxes are spaces where children play freely, where they can build, deconstruct, and start over without the constraints of permanence. Sand is a malleable medium that allows for a multitude of shapes, and it can always return to being what it is: sand. Sandboxes – as instruments to fostering innovation – serve a very similar function. They create a bounded space akin to the "box" in the "sandbox" that allows the testing of new technologies and novel business models. This is done without necessarily compromising the whole, limiting systemic impacts, and allowing for a fresh start if necessary. As such, they provide an opportunity to propose an innovative tool in a controlled and supervised environment. In this regard, their primary goal is to enable the development of technology in a safe and controlled manner.[724]

In this way, they assist in conducting independent tests before a tool is put into practice, enabling the development of standards, principles, and methodologies. These testing environments can be

[724] Prevelakis, V., & Spinellis, D. (2001, June). Sandboxing Applications. In *Usenix annual technical conference, freenix track* (pp. 119-126). <https://www2.dmst.aueb.gr/dds/pubs/conf/2001-Freenix-Sandbox/html/sandbox32final.pdf>.

regulatory or operational. The latter was the case with the Sandbox developed within the framework of the “Data for Justice” project, conducted in partnership between ITS Rio and the Public Defender’s Office of the State of Rio de Janeiro (DPRJ).

Therefore, Sandboxes are important methodological instruments for the development of various technologies. The following chapter aims to present how this methodology works, the possible formats, and the format developed in the case study presented project, with the objective of demonstrating possible use cases and challenges in practice for future Sandboxes.

## 17.4 Distinguishing Regulatory and Operational Sandboxes

There are at least two major types of Sandboxes: i) operational ones, which aim to foster the development of new technologies and tools; and ii) regulatory ones, which allow for the analysis of the impacts of regulation on a specific technology or of a new technology or business model within an existing regulation. Refer to the table below to understand how each model works:Table 01: Distinguishing Operational and Regulatory AI Sandboxes

| | Regulatory Sandboxes | Operational Sandboxes |
|---|---|---|
| **Definition** | Embody a spirit of controlled experimentation. However, the focus is not solely on the technology itself, but on the regulation. This is due to the novelty of the untested new technological tools or business models and their impact. They may have impacts that are: i) *systemic,* where regulation might prove insufficient; or, even if localised, ii) *unclear,* in need of interpretative clarity. The regulatory context highlights the essential participation of an authority or a public agency with competence in the area being tested. | Focus on technology and the changes it can bring to the systems where the technologies might be implemented. Thus, they act as “testing grounds,” spaces for experimentation. It’s somewhat like a company with multiple branches setting aside one branch to test a new technology, such as a new payment system, changes in checkout procedures, or a different way of arranging products. This is done in a limited manner (perhaps just one branch) to practically assess the opportunities and challenges that might arise when the change is fully implemented. This creates a phase during which the envisioned model can be practically developed. As a consequence, it presupposes continuous analysis and the possibility of course corrections and feature adjustments. |

| | Regulatory Sandboxes | Operational Sandboxes |
|---|---|---|
| **Seek to answer the following questions** | What is the best regulation? What impacts can it have on existing regulation? Are new guidelines or norms necessary? | How to develop a specific technology responsibly? What are the impacts? What measures can be taken to prevent and mitigate these impacts? |
| **Examples** | Fintech-related regulatory sandboxes such as those maintained by the UK Financial Conduct Authority (FCA) and Brazil's Central Bank (BACEN).[725]<br>Also, there are Regulatory sandboxes in artificial intelligence such as the EU AI regulatory sandbox pilot program in Spain.[726] | Operational Sandboxes are commonplace in the software development industry to debug and test software[727] and also in cybersecurity for malware detection.[728] |

Table 1: Distinguishing Operational and Regulatory AI Sandboxes

In the project with DPRJ, an Operational Sandbox was developed. To achieve this, participation dynamics were established through the formation of a Multistakeholder Committee to assess and contribute to the development of the Operational AI Sandbox. The involvement of diverse stakeholders was crucial to understand risks, principles, and limitations for developers, defenders, officials, and citizens, who were either involved in or affected by the development of this technology. Hence, discussions were held on the impacts and ways to mitigate the risks associated with the use of AI technology, especially in projects involving human rights.

## 17.5 Fostering multistakeholder social participation for the development of AI projects.

In order to build ethical and responsible AI, there is a need for both representation and participation from a wide range of sectors of society that may be impacted by the technology under design. Thus, there should be engagement of various stakeholders, encompassing groups and individuals from civil society, academia

[725] <https://patents.google.com/patent/US8799862B2/en>.

[726] <https://ieeexplore.ieee.org/abstract/document/5665792>.

[727] Learn more about this kind of regulatory sandboxes on this article by the World Bank:<https://blogs.worldbank.org/psd/four-years-and-counting-what-weve-learned-regulatory-sandboxes> (Access on September 12, 2023).

[728] Learn more about AI regulatory sandboxes by accessing this OECD report on the subject:<https://www.oecd.org/sti/regulatory-sandboxes-in-artificial-intelligence-8f80a0e6-en.htm> (Access on September 12, 2023).

and private sector -besides competent public bodies. Participation in technology projects contribute to the establishment of values and principles aligned with human rights and fundamental freedoms – especially the rights of marginalised communities, or those in a situation of vulnerability -, labour rights, environmental preservation, ecosystems, as well as ethical and social implications.[729] Engaging stakeholders through various forms of social participation can serve as a central means for materialising the values and principles to be adopted. Through the collection of information – whether through workshops, surveys, or in-depth interviews – concrete actions can be made feasible for adoption in technology development.

Furthermore, involving groups from different sectors, such as multistakeholder Committees, enables the assessment of the impacts of AI systems and how this technology might affect various groups and individuals.

## 17.6 Multistakeholder Committees

Considering the significance of a multistakeholder approach to technological development, cooperation among different sectors and stakeholders through Multistakeholder Committees can be the key to mitigating many of the risks and negative technological impacts.

This institution, which can be temporary or permanent, is tasked with guiding and advising on actions of transparency and social participation in the development of ethical and responsible AI. Members of the Committee are in charge of assisting in decision-making to align the development of AI tools with principles of human rights and fundamental freedoms, ensuring that project activities become more transparent and garner greater engagement from involved stakeholders.

It is possible to establish Thematic Subcommittees for the execution of specific activities, in which the participation of other representatives is a matter of free choice. These arrangements are an important instrument for addressing subjects that involve various sectors, bodies, or departments. Additionally, they promote the creation

[729] Leslie, D., & Briggs, M. (2021, March 20). Explaining decisions made with AI: A workbook (Use case 1: AI-assisted recruitment tool). ArXiv.org. <https://arxiv.org/abs/2104.03906>.

of task groups to collaborate on discussions regarding risks, opportunities, and potential uses of the technology. There can be thematic committees or meetings which can focus their efforts on particular issues such as data protection, IT infrastructure, UX (user experience), etc.

In the project with the Public Defender's Office of the State of Rio de Janeiro, a Multistakeholder Committee, composed of members from civil society, academia, private sector and government, supported the collaborative design of the AI tool throughout its full development under the Operational Sandbox. The Committee included experts from the Public Defender's Office of the State of Rio de Janeiro, the NGOs PretaLab and Institute for Health Policy Studies (IEPS), Oswaldo Cruz Foundation (FIOCRUZ), the National School of Public Health Sérgio Arouca, and two AI consulting and software development startups, ASK and Hacklab.

The collaboration with these specialists allowed the project to address significant issues of the tool development and anticipate potential challenges. Their roles included defining the principles which guided both the Sandbox and actual development of the concept and implementation of the AI tool. The diverse knowledge and varied experience of the members were of significant value to the results of the project. As a first step, it was necessary to sensitise the Committee members involved and train them on the subject under analysis, in this case, access to health care and the work of the public defenders in achieving that. This was done so that all participants could have a thorough understanding of the topic and enable them to enhance their active participation in the discussions that took place throughout the project.

Next, the Committee worked to determine what question this AI Sandbox could answer and how it could support the activities of the public servants from DPRJ, with the goal of promoting greater efficiency and effectiveness in upholding the human rights of those assisted. Subsequently, the principles of AI that could be relevant and important to the group were defined. These principles were incorporated and applied to the development of the proposed technology in the project. One clear example of this was the principle of personal data protection where due to the suggestion of participants, it was decided that data referred to the tool would be anonymised under techniques of statistical anonymization.

Furthermore, the management and prioritisation of information carried out with the DPRJ facilitated the focusing of efforts on specific themes. This evaluation helped assess the availability of databases and how they could be utilised for the construction of the AI.

In order to structure the Proof of Concept (PoC) for the technology, the necessary elements for its development were validated. This included information and available databases, assessing whether they would address the question and the problem that the technology aimed to solve. Other steps involved testing and validating the Proof of Concept in order to be able to develop the AI tool itself, as well as dealing with complexities around timeframe and challenges posed by the structure, anonymization and extraction of necessary data.

## 17.7 Analysis of ethical issues related to AI through Multistakeholder Committees

Developing an AI system without considering its potential transformative and long-term effects on individuals and society can lead to (re)production of discrimination and social inequality.[730] To ensure that the implementation of an AI system remains sound and supports the sustainability of the communities it affects, developers are recommended to proceed with ongoing sensitivity to the actual impacts their system may have.[731] In this context, Risk Analysis is a necessary component of a sandbox model for the development of technologies that utilise AI models, to determine the ethical permissibility of the project. It's recommended that the analysis of (positive and negative) impacts from the production software development occurs in two stages:

[730] NOBLE, Safiya. Algorithms of oppression: How search engines reinforce racism. NYU Press; BROWNE, Simone. Dark matters: on the surveillance of blackness. Durham: Duke University Press, 2015.

[731] Leslie, D. (2019). Understanding artificial intelligence ethics and safety A guide for the responsible design and implementation of AI systems in the public sector Dr David Leslie Public Policy Programme. Understanding Artificial Intelligence Ethics and Safety. <https://doi.org/10.5281/zenodo.3240529>.

Pre-existing Conditions

**a.** How was it done before?

Post-implementation

**b.** Can impacts be measured, such as:

**c.** Were distortions in AI responses reduced?

**d.** Were predicted effects mitigated?

This approach ensures that the ethical and societal implications of AI technologies are thoroughly evaluated and addressed, promoting fairness and minimising adverse consequences. The project team based their considerations on the possible societal effects of the project's AI system on ethical principles. The team engaged with the Multistakeholder Committee members to assess the social impact and sustainability of their AI project through a preliminary Risk Assessment. Conducting a prior impact analysis before the development of the tool, regardless of whether the AI is used for providing public services or in administrative capacities, aimed to instil confidence that the project and the implementation of the AI system by the public sector agency took ethical and responsible principles into account to promote human rights.

Furthermore, the participation of a diverse array of stakeholders in this process illuminated invisible risks that could potentially affect individuals and the public good. This approach also endorsed transparent innovation practices and well-informed decision-making. Examples of such risks refer to the potential exclusion of certain areas and individuals due to re-prioritization of resources. This was noted by the committee and addressed through rearranging the tool to include outliers.

The pre-assessment of potential risks and considerations about the design of AI for the DPRJ, as carried out during the Data for Justice project, through the Multistakeholder Committee, was divided into four parts, as described in Table 02. The intention was to clarify the ultimate purpose of the AI, identify potential areas of action and impact markers, and then develop risk mitigation strategies throughout the entire lifecycle of the tool.

| Pre-Assessment of AI's Potential and Risks | |
|---|---|
| **Question** | **Points Evaluated** |
| 1- Which tool would we like to have and which one do we have? | • Most Requested Items in terms of Medication;<br>• Diagnosis of a social (health) reality;<br>• Understanding the effectiveness of judicial intervention in terms of medication/treatment requisition;<br>• Scalability of the technology;<br>• Preserved procedural guarantees;<br>• Understanding the "map" of health needs not automatically met by current policies;<br>• Preserved procedural guarantees;<br>• Innovation;<br>• Understanding the personal situation of beneficiaries[732] at a macro level. |
| 2- What factors can this tool impact? | • Possible diversion of the DPRJ's attention to specific issues rather than the whole;<br>• Does it affect the DPRJ's relationship with the judiciary?<br>• How to generate insights for the whole based on specific problems;<br>• Does it affect the speed of decision-making?<br>• Understanding of the tool by the employees;<br>• Visualisation of demand vs. necessity/priority;<br>• Does it require the digitization and standardisation of any work process?<br>• Difficulty in accessing data impacting development;<br>• What are we calling innovation? How do we measure the impact?<br>• Does AI technology promote innovation in the judiciary system and the ecosystem?<br>• Impact on the rational use of medication; |
| 3- What criteria can we use to analyse the impacts? | • Potential for discrimination, exclusion, or bias;<br>• Understanding of the tool by the employees;<br>• Inequity of judicialization;<br>• Speed of decision-making;<br>• Effective visualisation;<br>• Response time;<br>• Usability of the tool. |

[732] In the Brazilian Unified Health System (SUS), a beneficiary is a person entitled to healthcare services and assistance provided by the system. SUS is a public healthcare system that aims to ensure universal, equitable, and free access to healthcare for all Brazilian citizens. Beneficiaries of SUS include all residents in Brazil, whether they are Brazilian or foreign nationals, as well as individuals passing through the country. SUS is primarily funded by the federal, state, and municipal governments, and its operation is based on principles such as equity, comprehensiveness, and universality. Therefore, all citizens have the right to be beneficiaries of SUS, which means they can seek medical care, surgeries, medications, and other healthcare services, free of charge, when necessary, regardless of their economic, social, or health status.

| Pre-Assessment of AI's Potential and Risks | |
|---|---|
| 4- Can we mitigate potential negative impacts? | • Respecting the right to privacy governed by LGPD (Brazil's General Data Protection Law);<br>• Data anonymization;<br>• Security measures, such as encryption;<br>• Training administrative staff to use the tool;<br>• Indication that there might be demands not covered by the results provided by the developed technology;<br>• Capable staff who can explain and differentiate that the demand provided by the technology may not necessarily be a priority/need in public management. |

Table 02: Questions and points of consideration brought up by the multistakeholder Committee during the pre-assessment of potential risks and considerations about the design of AI for the DPRJ.

## 17.8 Ethical Principles of AI Development

The misuse, abuse, or inadequate design of AI tools can lead to individual and societal harm,[733] ranging from discrimination, non-transparency and unjustifiable outcomes to privacy infringements and exclusion. Therefore, it is recommended to pay attention to ethical considerations and establish policy foundations based on ethical principles. These principles play a crucial role in every AI project, particularly when developed within the justice ecosystem,[734] in order to mitigate risks and ensure reliable, secure, and high-quality outcomes.

Creating a project delivery environment that allows for ethical design and implementation of AI systems promoting human rights requires a multidisciplinary team effort. It demands the active cooperation of all team members, both in maintaining a culture of accountability and in executing a governance framework that embraces ethically sound practices across all stages of the innovation and implementation lifecycle.

The introduction of ethical principles in the reported project occurred through meetings with the Multistakeholder Committee, where the

[733] Leslie, D., 2019 (op. cit.).

[734] JUNQUILHO, Tainá Aguiar. Inteligência Artificial e Direito: limites éticos. Juspodivm: Salvador, 2022.

concepts of FAIR principles[735] and the UNESCO guidelines for creating ethical AI[736] were presented, discussed and positioned within the context of the project both in terms of theme, area of analysis and local social and cultural specificities. Table 3 provides an overview of the principles discussed:

| | |
|---|---|
| **UNESCO** | • Proportionality and Non-Harm<br>• Security and Protection<br>• Equity and Non-Discrimination<br>• Sustainability<br>• Right to Privacy and Data Protection<br>• Transparency and Explainability<br>• Responsibility and Accountability<br>• Awareness and Literacy<br>• Multistakeholder Governance and Collaboration<br>• Adaptive Promotion of Human Values |
| **FAIR** | • Findable<br>• Accessible<br>• Interoperable<br>• Interoperable |

Under a dynamic session conducted with the Committee participants, it was asked of them to:

- Identify key existing ethical principles.
- Consider additional principles not listed that should apply to the project at hand. (These new principles were categorised as “Innovation.”)
- Divide and categorise the principles into “Priorities” (essential to the project) and “Secondary” (less fitting in the project’s context).
- Propose actions to ensure that the principles are implemented throughout the project.
- In the priority category, the following principles were listed:
- Right to Privacy and Data Protection
- Adaptive Promotion of Human Values;
- Equity and Non-Discrimination;

[735] The Fair Principles first appeared in a 2016 publication written by a consortium of scientists and organisations and entitled “The FAIR Guiding Principles for scientific data management and stewardship”. For more information see: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4792175/> (Access on September 12, 2023). GO FAIR. (2017). FAIR Principles – GO FAIR. GO FAIR. <https://www.go-fair.org/fair-principles/>.

[736] UNESCO (2022). Recommendation on the ethics of artificial intelligence. Unesco.org. <https://unesdoc.unesco.org/ark:/48223/pf0000381137>.

- Awareness and Literacy;
- Security and Protection;
- Proportionality and Non-Harm;
- Responsibility and Accountability;
- Multistakeholder Governance and Collaboration;
- All of the FAIR principles;

In the "Innovation" category the principle of "Replicability" was added, whereas the principles of "Sustainability" and "Supervision and Human Determination".[737]

## 17.9 Governance challenges faced during the development of AI

Throughout the project, challenges emerged in terms of governance of the project. The role of the Multistakeholder Committee was essential in raising questions about which principles were to be prioritised, that is: what do we consider ethical AI? Authors from the Global Majority have been pointing to the need to contextualise international ethical AI guidelines,[738] largely developed in the Global North and deployed in other regions of the planet without caring for refining these models to the demographic and cultural reality of the communities of these regions. The Committee's meetings were crucial to understand that principles must be understood in their cultural, linguistic, geographic, and organisational context, and certain themes will be more relevant to a specific context and audience than others. Furthermore, the impact of these principles depends on their integration into a broader governance ecosystem, including laws, regulations, relevant policies (such as national AI plans), as well as professional practices and daily routines.

Most social participation processes come with challenges in engaging stakeholders. For the development of this project, involving all stakeholders was considered essential, including public sector

[737] This AI principle was not listed in either the FAIR principles or the UNESCO guidelines; it was conceived by the Committee during the working session. However, after discussion, it was moved from the "Innovation" category and placed in the secondary category.

[738] Sambasivan, N., Arnesen, E., Hutchinson, B., Doshi, T., & Prabhakaran, V. (2021). Re-imagining Algorithmic Fairness in India and Beyond. Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. <https://doi.org/10.1145/3442188.3445896>.

employees, legal professionals, legal tech companies, or scientists and researchers in the field of human rights in health and technology. This brought the challenge of computer literacy in conveying and training on the topic. In the specific case reported in this article, even though the majority of the Multistakeholder Committee and the members of the DPRJ knew some basic concepts of AI literacy, it was deemed necessary to hold training sessions, conducted by ITS Rio, on the ethical impacts and regulation regarding the adoption of AIs in the Brazilian public service.

This effort made it possible to convey the scope and potential impact of introducing artificial intelligence applications into the Brazilian judicial ecosystem—within the scope of the study case of the model designed for DPRJ—and to define the ethical framework in which it could operate.

During the data extraction process, the project had to address the need for anonymization of personal data of DPRJ beneficiaries present in the institution's database (named "Verde"). This required mapping the information to identify which tables could contain personal data and what forms of anonymization would be necessary.

This process of data extraction and anonymization had another layer of complexity due to the limited computational infrastructure at DPRJ. In fact, the DPRJ outsources much of its computational power to other entities, particularly the Federal University of Rio de Janeiro (UFRJ), with whom they maintain a technical cooperation agreement. Hence, this other entity had to be involved in the process of extracting and anonymizing the data from their database. Similarly, involving UFRJ or another external entity will be necessary if the cooperation between DPRJ and the university ends in the future, in any process involving the development of new AI technologies. Governance issues of the whole process had to be addressed. The Multistakeholder Committee with its specificities made it easier to bring the necessary new player into the fold. It had at its core a principle of participation and inclusion that allowed for adequate circulation of information and easy access to solutions.

## 17.10 Final Considerations

The development and study of Artificial Intelligence, like that of any complex software, is inherently a sociotechnical endeavour. This means recognizing the interconnection, indivisibility, and indeterminacy of the technical and the social aspects. The sociotechnical perspective understands them as mutually determining, rather than artificially separating them a priori.[739] Therefore, a comprehensive study of the reconfiguration in the relationship between human and computational actors at DPRJ would require a period of observation and deep immersion.

Ethical considerations for an AI, as mentioned earlier, cannot be simply implanted from ethical models developed in the Global North and expected to fit the context of AI development in the Global Majority, especially within DPRJ. On the contrary, it requires an "anthropophagic" process (digestion and adaptation for incorporation) on the part of DPRJ to absorb and reconstruct any set of ethical principles, taking into account its local specificities. This process will require observing the use of AI by DPRJ and how the information provided by it modifies the work of Defenders. Based on these observations and on the dense description of the changes brought about by the introduction of AI, the Multistakeholder Committee can be utilised as a space for constructing an ethical framework tailored to the context of DPRJ.

Similarly, regarding the integration of a new computational tool into the IT structure of DPRJ (such as an AI), can only be implemented after observing and adapting the current IT governance practices to provide maintenance and improvements to the AI tools developed, making them compatible with the "Verde system". Again, the Multistakeholder Committee can be explored as a space for constructing these new governance processes.

Lastly, a significant takeaway from this project is the realisation that, for implementation in other Public Defender's Offices, it is recommended to conduct prior mapping of the actors involved in

[739] Cukierman, H. L., Teixeira, C., & Prikladnicki, R. (2007). Um olhar sociotécnico sobre a engenharia de software. *Revista de Informática Teórica e Aplicada*, *14*(2), 199-219. <https://doi.org/10.22456/2175-2745.5696>.

the internal IT governance of these institutions. Additionally, it was discovered that, in the specific context of this project, the creation of processes and documentation for anonymizing beneficiary data and managing internal computational systems at DPRJ was necessary.

# 18 Regulatory Sandboxes as Tools for Ethical and Responsible Innovation of Artificial Intelligence and their Synergies with Responsive Regulation

**Thiago Moraes, Vrije Universiteit Brussels (VUB) and Universidade de Brasilia (UnB)**

## Abstract

This paper explores the role of regulatory sandboxes as tools to foster ethical and responsible innovation in artificial intelligence (AI) systems and discusses the synergies of sandboxes with responsive regulatory theory. The analysis is carried out through bibliographical research with focus on experiences from the Global South (Brazil, Colombia and Singapore) and European countries. To argue about the importance of sandboxes as drivers for ethical innovation in AI, the study (i) starts from sandboxes based on sectoral regulatory licensing regimes, such as the financial and telecommunications sectors, (ii) advances to the experiments carried out in regulatory regimes based on risk and fundamental rights protection, such as personal data protection, and (iii) analyses the legislative debates on regulatory sandboxes in the contexts of AI regulation in the European Union and in Brazil, in order to reflect which of the previous approaches AI sandboxes are closest to. Finally, (iv) the study reflects on the synergies of sandboxes with the theory of responsive regulation, so that they can be integrated into regulatory strategies which adopt this theory.

## 18.1 Introduction

The growing use and development of artificial intelligence (AI)[740] has promoted a global race for regulatory frameworks, with the goal of

[740] According to the Organization for Economic Co-operation and Development (OECD), an Artificial Intelligence (AI) System is "a machine-based system that can, for a given set of human-defined goals, make predictions, recommendations or decisions that influence real or virtual environments." The definition has been adopted in different legal systems, such as the European, in the context of the AI Act and in Brazil (Bill n. 2338/2023). This OECD definition is currently undergoing an update process and it is expected that a new version of the definition will soon be presented. Hersey, F. (2023). *EU AI Act definition of AI aligns with OECD definition, biometric risk updated,* Biometric Update.com. Retrieved from: <https://www.biometricupdate.com/202303/eu-ai-act-definition-of-ai-aligns-with-oecd-definition-biometric-risk-updated>.

developing so-called "trustworthy" AI systems.[741] Regardless of the region, a concern continually raised by the economic sector is the risk that (over)regulation will stifle the development of innovative solutions.

This concern is not new, and David Collingridge identified it in 1980 as the control dilemma.[742] The regulator, in its role, wants technology to be better controlled to avoid harmful social consequences. However, he faces a double problem: on the one hand, there is an information problem, since such damage can only be accurately predicted when the technology is more widely developed and widely used. In another, there is the problem of power, because as technology becomes intertwined with society, it decreases the ease of influencing its social, political and innovation trajectories.[743] This paradox between regulation and technology is known, in homage to its author, as the Collingridge Dilemma.

While Collingridge may be criticized for having an "anti-innovation" perspective, since it could incentivize conservative approaches to inhibit innovation before it develops, his work has actually deeply contributed to the field of responsible research and innovation (RRI), and several of its core characteristics, such as a focus on addressing significant socio-ecological needs and challenges; a commitment to actively engage a range of stakeholders for the purpose of substantively improving decision-making and mutual learning; a dedicated attempt to anticipate potential problems; and a willingness among all participants to act and adapt according to these ideas.[744] Collingridge's approach emphasises active processes

[741] Smuha, N. A. (2021). From a 'Race to AI' to a 'Race to AI Regulation' – Regulatory Competition for Artificial Intelligence. *International Journal of Law, Innovation and Technology,* 13. Retrieved from: <https://ssrn.com/abstract=3501410>.

[742] Collingridge, D. (1980). *The Social Control of Technology*. Frances Pinter Publisher Ltd, 16.

[743] A current example of this is the case of social networking platforms. As they grew and became intertwined in our society, it becomes increasingly challenging for the State to stand up to the technology companies that control these environments, creating a point of friction between the State and the economic sector. In the Brazilian context, when the National Congress proposed the Bill n. 2630/2023, to create stricter rules of transparency and fight against misinformation for digital platforms, companies like Google and Facebook reacted by promoting biased campaigns, which resulted in the suspension of the legislative proposal. Weterman, D.; Affonso, J. (2023). *Pressão e ameaça no Congresso: como Google e Facebook derrubaram o PL 2630 das Fake News em 14 dias.* São Paulo: Estadão. Retrieved from: <https://www.estadao.com.br/politica/pressao-e-ameaca-no-congresso-como-o-google-derrubou-o-pl-2630-das-fake-news-em-14-dias/>.

[744] Gennus, A.; Stirling, A. (2017). Collingridge and the dilemma of control: Towards responsible and accountable innovation. *Research Policy*. Retrieved from: <http://dx.doi.org/10.1016/j.respol.2017.09.012>.

of learning from a particular class of past decisions in order to inform future de cision-making about technology development, scientific research and innovation, being pragmatically concerned with the qualities of emerging innovations, rather than consequentially with their outcomes.

In the context of data-driven technologies such as Artificial Intelligence, Collingridge's dilemma is increasing due to rapid technological developments which require regulators to make decisions in the absence of reliable risk information or prior knowledge of technological developments.[745] To fill this ever-growing gap, a proposed solution is the development of new regulatory approaches that focus on public policy experimentation.[746]

One of these approaches that has aroused growing interest (or curiosity) on the part of regulators are regulatory sandboxes, collaborations that bring together regulators and organizations that develop new technologies and processes to test innovations in relation to the regulatory framework.[747] The growing interest in the topic has led legislators to include regulatory sandbox provisions in AI legislative proposals, such as the European Union (EU) AI Act and the replacement text of the Brazilian AI Bill.

There are several experiments with regulatory sandboxes that regulators are implementing. Originally, sandboxes emerged in the financial sector in the mid-2010s, when new services in this market began to use emerging technologies: fintechs.[748] In this context, the main goal of sandboxes was to allow regulatory flexibility, temporarily suspending rules in this sector to reduce regulatory barriers for entrants and allow the regulator to better understand what are the benefits brought by these new services.

[745] Bromberg, L; Godwin A; Ramsay, I. (2017). Fintech sandboxes: Achieving a balance between regulation and innovation. *Journal of Banking and Finance Law and Practice*, 28(4). Retrieved from: <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3090844>.

[746] Vermeulen, E.; Fenwick, M.; Kaal W. A. (2017). Regulation Tomorrow: What Happens when Technology is Faster than the Law? *American University Business Law Review*, 6(4). Retrieved from: <http://dx.doi.org/10.2139/ssrn.2834531>.

[747] The Datasphere Initiative. (2022). *Sandboxes for data: creating spaces for agile solutions across borders*. Retrieved from: <https://www.thedatasphere.org/datasphere-publish/sandboxes-for-data/>.

[748] World Bank. (2020) *Global Experiences from Regulatory Sandboxes*. Washington, DC: World Bank, 2020. Retrieved from: <https://openknowledge.worldbank.org/handle/10986/34789>.

Over time, other regulators began to use the experimental regulation promoted by sandboxes in their own contexts. In Brazil,[749] for example, in addition to the financial sector,[750] it is possible to identify initiatives in the telecommunications sector,[751] in health,[752] and in transport infrastructure.[753] In all these cases, the approach involves the lifting of barriers and regulatory simplification to foster innovation in the respective sectors.

However, sandboxes have been used by some regulatory authorities with different goals. Data Protection Authorities (DPAs) oversee regulatory frameworks which are not based on prior authorizations (hereinafter, licensing regimes),[754] but on the risk-based approach, in which regulatees don't need licenses to operate in the market but must proactively demonstrate compliance to the regulator. For example, in the Brazilian data protection legislation, Law n.

[749] In addition to the national initiatives mentioned in this paragraph, there are also regional initiatives in Brazil, such as in the cities of Rio de Janeiro (RIO DE JANEIRO, 2022) and Foz do Iguaçu (FOZ DO IGUAÇU, 2020). Due to limitations of scope, this paper will not focus on municipal cases. Foz do Iguaçu (2020), *Decreto nº 28.244, de 23 de Junho de 2020,* Regulamenta no âmbito do Município de Foz do Iguaçu, a instituição de ambientes experimentais de inovação científica, tecnológica e empreendedora, sob o formato de Bancos de Testes Regulatórios e Tecnológicos – "Programa Sandbox – Foz do Iguaçu". Retrieved from: <https://leismunicipais.com.br/a/pr/f/foz-do-iguacu/decreto/2020/2825/28244/decreto-n-28244-2020-regulamenta-no-mbito-do-municipio-de-foz-do-iguacu-a-instituicao-de-ambientes-experimentais-de-inovacao-cientifica-tecnologica-e-empreendedora-sob-o-formato-de-bancos-de-testes-regulatorios-e-tecnologicos-programa-sandbox-foz-do-iguacu>. RIO DE JANEIRO (2022) Sandbox.rio. Retrieved from: <https://www.sandboxrio.com.br/sobre.html>.

[750] There are sandbox programs being conducted by three authorities of the National Financial System. For more details, see Part I of this article.

[751] The National Telecommunications Agency – ANATEL, conducted a public consultation (n. 41, of June 9, 2022) for simplification on the regulation of telecommunications services, having as one of its premises the constitution of a regulatory sandbox program, entitled "experimental regulatory environment". Agência Nacional de Telecomunicações – ANATEL. (2022). *Consulta pública nº 41,* Brasília: ANATEL. 2022. Retrieved from: <https://apps.anatel.gov.br/ParticipaAnatel/VisualizarTextoConsulta.aspx?TelaDeOrigem=3&ConsultaId=10021>.

[752] In October 2022, the National Health Agency – ANS, held a webinar to discuss proposals for experimental regulation such as the sandbox, together with the regulated sector. Taking subsidies was a first step towards developing the initiative. Agência Nacional de Saúde Suplementar – ANS. (2022). *ANS promove webinar sobre SandBox Regulatório Prudencial*, Brasília: ANS. Retrieved from: <https://www.gov.br/ans/pt-br/assuntos/noticias/periodo-eleitoral/ans-promove-webinar-sobre-sandbox-regulatorio-na-saude-suplementar>.

[753] The National Land Transport Agency – ANTT, established the rules for the constitution and operation of the Agency's regulatory sandbox with Resolution n. 5,999, of 2022, also considered as an experimental regulatory environment. ANTT adopted a regulatory sandbox model that allows testing innovative products or services and innovative regulatory solutions. Agência Nacional De Transportes Terrestres – ANTT. (2022). *Resolução nº 5.999, de 3 de novembro de 2022.* Brasília: ANTT. Retrieved from: <https://www.in.gov.br/en/web/dou/-/resolucao-n-5.999-de-3-de-novembro-de-2022-441284496>.

[754] According to Di Pietro (2020), licensing regimes to perform a public service is typical of the regulatory agencies of Brazilian administrative law, which, like BACEN and ANATEL, regulate and control the activities that constitute the object of authorization. Di Pietro, M. S. Z. *Direito administrativo 3 ed.* Rio de Janeiro: Forense, 2020.

13.709/2018 (LGPD), the accountability principle[755] is one of its key elements to promote risk-based regulation.

Thus, DPAs' sandbox programs have focused on promoting the implementation of data protection principles and of privacy by design.[756] The proposals implemented by these authorities are focused on promoting responsible innovation, in line with the respective data protection legislation.

When advancing to the debate on the use of regulatory sandboxes in the context of artificial intelligence regulations, which are still under development,[757] some questions arise: should AI sandboxes follow the financial sector approach, in focusing on waivers to reduce barriers to innovation? Or should they focus on fostering responsible innovation, like the experiments being conducted by DPAs?

This paper intends to explore these questions based on the analysis of sectoral sandboxes, based on authorization regimes, and risk-based sandboxes, such as data protection legal frameworks. The analysis is carried out through a bibliographical survey of reports from public and private, national and international organizations on regulatory sandboxes and academic paper on the theory of responsive regulation. A current challenge is that there is little academic literature on sandboxes, perhaps due to the novelty of the topic.

This study also intends to briefly reflect upon the role of sandboxes, as regulatory tools for regulators' strategies. To this end, it will present possible relationships between regulatory sandboxes and the theory of responsive regulation. This theory reverses the traditional structure of regulatory enforcement, of command and control, to one in which the authority, through dialogical proceedings, flows and definition of competences, saves on coercive means, in favour of collaboration for the promotion of virtuous behavior of the regulatee.[758] However,

[755] LGPD, Art. 6, X – accountability: demonstration, by the agent, of the adoption of effective measures capable of proving compliance with personal data protection rules, and the effectiveness of these measures (author's translation).

[756] This concept will be explained in Part II of this study.

[757] At the time of writing this paper, no proposal for a law has been identified in a national or international environment that regulates artificial intelligence regulatory sandboxes. The most advanced proposal in this regard is that of the European Union, which will be discussed in more detail in Part III.

[758] Aranha, M. I., & Lopes, O. A. (2019). (rep.). *Estudo sobre Teorias Jurídicas da Regulação apoiadas em incentivos. Pesquisa e Inovação Acadêmica em Regulação apoiada em incentivos na Fiscalização Regulatória. Projeto ANATEL-UnB (Meta 5).* Brasilia: Centro de Políticas, Direito, Economia e Tecnologias das Comunicações da UnB.

given the paper's limitations , this analysis serves to start a debate which could be further developed in future studies.

## 18.2 Sandboxes Based On Sectorial Licensing Regimes

As literature suggests, regulatory sandboxes began to be used in the context of the financial sector, when faced with new market actors that introduced innovations in financial markets through the intensive use of information technology, potentially creating new business models – fintechs. According to the Central Bank of Brazil – BACEN, in Brazil, there are several categories of fintechs: credit, payment, financial management, loan, investment, financing, insurance, debt negotiation, exchange and multiservice.[759]

In the international context, the pioneer in the application of sandboxes was the British agency Financial Conduct Authority (FCA), in 2016.[760] The main goal was to test new and innovative financial services without incurring all the normal regulatory consequences of engaging in these activities. The idea spread worldwide, and in 2020, a World Bank report identified that 57 countries operated 73 fintech sandboxes. Brazil is in this list: in 2019, the Securities and Exchange Commission (CVM), the Private Insurance Superintendence (Susep) and BACEN jointly published a statement on the implementation of the regulatory sandbox in the respective markets of operation.[761] According to the institutions, their goals are to ensure innovation and business diversity, promoting competition and meeting user needs.

Although the financial sector is the "birthplace" of sandboxes, the tool has spread to other sectors. In 2019, the German government published a specific study on regulatory sandboxes in which, in addition to presenting characteristics of the concept and good practices for their use, it shared 27 experiences carried out in several sectors,

[759] Banco Central do Brasil – BACEN. (2020). *Fintechs*. Brasília: BACEN. Retrieved from: <https://www.bcb.gov.br/estabilidadefinanceira/fintechs>.

[760] World Bank. (2020) *Global Experiences from Regulatory Sandboxes*. Washington, DC: World Bank, 2020. Retrieved from: <https://openknowledge.worldbank.org/handle/10986/34789>.

[761] Organization for Economic Cooperation and Development – OECD. (2018). *Financial Markets, Insurance and Private Pensions: Digitalisation and Finance*. Paris: OECD. Retrieved from: <https://www.oecd.org/finance/private-pensions/Financial-markets-insurance-pensions-digitalisation-and-finance.pdf>.

including energy, transport and logistics infrastructure.[762] According to them, regulatory sandboxes have three main characteristics: (i) they are test zones established for a limited time, covering a specific sector, in which innovative technologies and business models can be experimented and made available to the public; (ii) they depend on regulatory flexibility or a regulation in which there is no immediate sanction for not strictly complying with a rule; (iii) they imply an interest in regulatory discovery, allowing the regulator to learn for the development of future norms and public policies.

Although methodologies may vary according to the regulatory authority and the sandbox's goals, according to European financial sector authorities, in general, a sandbox consists of the following steps[763]: (i) proposal submission and evaluation; (ii) preparation of selected participants; (iii) testing and monitoring of initiatives; (iv) evaluation of experience and exit. The process is structured in such a way as to guarantee participation's isonomy and exchange of knowledge among the actors involved.

As mentioned in the introduction, there are sandbox initiatives being developed in Brazil in several regulated sectors. What is observed in these programs is an interest in regulatory simplification during experimentation, allowing regulators to develop flexible licensing regimes for testing and to be able to develop future policies prototyping. Thus, the three characteristics identified by the German government may also apply to Brazilian cases.

To ensure legal certainty in the implementation of regulatory sandboxes in Brazil, Legal Startups Framework, Complementary Law No. 182/2021,[764] brought, in its art. 11, that "bodies and entities of the public administration with competence for sectoral regulation may, individually or in collaboration, within the scope of experimental regulatory environment programs (regulatory sandbox), remove the

[762] BMWi – Federal Ministry for Economic Affairs and Energy. (2021). (rep.). Making space for innovation: The handbook for regulatory sandboxes. Retrieved from <https://www.bmwk.de/Redaktion/EN/Publikationen/Digitale-Welt/handbook-regulatory-sandboxes.pdf%3F__blob%3DpublicationFile%26v%3D2>.

[763] European Securities and Markets Authority – ESMA; European Bank Authority – EBA; European Insurance And Occupational Pensions Authority – EIOPA. (2018). *FinTech: Regulatory Sandboxes and Innovation hubs*. Brussels: ESMA, EIBA & EIOPA. Retrieved from: <https://www.esma.europa.eu/sites/default/files/library/jc_2018_74_joint_report_on_regulatory_sandboxes_and_innovation_hubs.pdf>.

[764] Brazil. (2021). *Lei Complementar nº 182/2021, de 1º de junho de 2021.* Marco Legal das Startups. Retrieved from: <https://www.planalto.gov.br/ccivil_03/LEIS/LCP/Lcp182.htm>.

incidence of norms under their competence in relation to the regulated entity or to groups of regulated entities."[765] For example, in ANATEL's proposed sandbox,[766] participants must obtain authorization for testing, a much simpler procedure than what an official license would be. Once the testing period is over, ANATEL may allow interested companies to operate the innovative business model tested in the sandbox, while it proceeds with the updating of relevant norms.

It is thus evident that sandbox programs such as those of ANATEL and BACEN make sense in regulated sectors where license to operate is required, reducing regulatory barriers to foster innovation. However, this paper's author believes that such an approach does not appear to be ideal in regulatory regimes that are not based on licensing, such as the risk-based approaches of data protection legislations. In the next part of this article, the experimental regulation initiatives of authorities in this other environment are analysed.

## 18.3 Sandboxes In Risk-Based Regulatory Regimes – The Experiences Of Data Protection Authorities

It's not just sectoral regulators who have been interested in sandboxes. Since FCAs' initiative, experimentation has happened in regulatory contexts not based on licensing regimes, such as those of personal data protection regulation. It is possible to highlight initiatives implemented by DPAs, such as Singapore's Personal Data Protection Commissioner – PDPC (2017)[767]; the United Kingdom's Information Commissioner's Office – ICO (2019)[768]; Norway's Datatilsynet (2021)[769]; Colombia's *Superintendencia de Industria y Comercio* – SIC, (2021)[770]; and France's

[765] Author's translation.

[766] ANATEL's proposed sandbox was presented on items 369 to 372 of public consultation No. 41 of 2022 (ANATEL, 2022).

[767] Personal Data Protection Commissioner – PDPC. (2017). *A Trusted Ecosystem for Data Innovation* Singapore: PDPC. Retrieved from: <https://www.pdpc.gov.sg/-/media/Files/PDPC/New_DPO_Connect/aug_2017/pdf/ATrustedEcosystemForDataInnovation.pdf.

[768] Information Commissioner's Office – ICO. (2019). Regulatory Sandbox, London: ICO, 2019. Retrieved from: <https://ico.org.uk/sandbox>.

[769] Datatilsynet. (2021). Sandbox for responsible artificial intelligence, Oslo: Datatilsynet. Retrieved from: <https://www.datatilsynet.no/en/regulations-and-tools/sandbox-for-artificial-intelligence/>.

[770] Superintendencia De Industria Y Comercio – SIC (2021). *Sandbox on privacy by design and by default in Artificial Intelligence*, Bogotá: SIC. Retrieved from: <https://www.sic.gov.co/sites/default/files/files/2021/ 150421%20Sandbox%20on%20privacy%20by%20design%20and%20by%20default%20in%20AI%20projects.pdf>.

*Commission Nationale de l'Informatique et des Libertés* – CNIL (2022).[771] Also, in Brazil, the *Autoridade Nacional de Proteção de Dados* (ANPD), informed, in May 2023, that it has started a technical cooperation with the Development Bank of Latin America – CAF, to develop a regulatory experimentation tool to foster innovation related to artificial intelligence (AI) under the scope of the LGPD.[772] The goals of the program are to allow participants to develop technologies that comply with personal data protection regulations, to be tested and analyzed in controlled environments, and that good practices are adopted to ensure compliance with personal data protection regulations.

Data protection legislations are usually structured by the risk-based approach, a partial meta-regulation mechanism, in which the personal data processing agent performs risk management assessments to verify its compliance with the legal rules of data protection, observing the obligations established according to the risk level of personal data processing activities.[773]

A concept directly related to the risk-based approach is privacy by design, a framework devised by Ann Cavoukian, which prescribes that privacy must be built directly into the design and operation of information technologies, business practices and network infrastructures.[774] Thus, it relates to the idea that data controllers should proactively incorporate personal data protection requirements into the entire lifecycle of processing personal information, from data collection to its erasure[775] (AEPD, 2019).

Privacy by design methodologies include the incorporation of an ethical dimension in the development of products and services and are related to the creation of technological measures to guarantee

[771] Commission Nationale de l'Informatique et des Libertés – CNIL (2022*). EdTech "sandbox": the CNIL supports ten innovative projects.* Paris: CNIL. Retrieved from: <https://www.cnil.fr/en/edtech-sandbox-cnil-supports-10-innovative-projects.

[772] ANPD. *ANPD formaliza cooperação técnica com o Banco de Desenvolvimento da América Latina – CAF*. 2023. Retrieved from: <https://www.gov.br/anpd/pt-br/assuntos/noticias/anpd-formaliza-cooperacao-tecnica-com-o-banco-de-desenvolvimento-da-america-latina-2013-caf>.

[773] Gellert R. (2020). The Risk-Based Approach To Data Protection, *Oxford Scholarship Online*, Oxford: Oxford University Press. Retrieved from: <https://academic.oup.com/book/40487.

[774] Information Privacy Commissioner – IPC. (2018). *Privacy by Design,* Toronto: IPC. Retrieved from <https://www.ipc.on.ca/wp-content/uploads/2018/01/pbd.pdf>.

[775] Agencia Española De Protección De Datos – AEPD. (2019). *Guía De Privacidad Desde El Diseño,* Madrid: AEPD. Retrieved from: <http://www.aepd.es/sites/default/files/2019-11/guia-privacidad-desde-diseno.pdf>.

the privacy and protection of personal data.[776] As examples of legislation that adopted versions of this concept, it is possible to mention the EU's General Regulation for the Protection of Personal Data – GDPR,[777] and the Brazilian LGPD.[778]

Privacy by design is highly compatible with DPAs' sandbox programmes. CNIL's President Marie-Laure Denis explained that promoting privacy by design was one of its sandbox goals, by integrating privacy protection concerns during the systems testing. SIC's sandbox pilot was entitled "Sandbox on privacy from conception and by default in Artificial Intelligence projects". PDPC started in 2022 a sandbox focused on privacy enhancing technologies (PETs), a set of technologies related to the concept of privacy by design. Thus, whether explicitly or implicitly, DPA sandboxes seem to be associated with the promotion of privacy by design, which, in turn, promotes the implementation of the principles inherent in data protection legislation.[779]

Thus, a switch of goals can be perceived. DPAs are not focusing on lowering regulatory barriers and providing temporary authorizations for innovators, not least because their regulatory regimes are not license-based. The goal of their programs is to foster ethical and responsible innovation, in compliance with data protection legislation and respecting data subject rights. While data protection regulations adopt the risk-based approach, they are also regimes for the protection of fundamental rights.

[776] Moraes, T. *et al.* (2021). Open data on the covid-19 pandemic: anonymisation as a technical solution for transparency, privacy, and data protection, International Data Privacy Law, 11(1), 32-47. Retrieved from: <http://dx.doi.org/10.1093/idpl/ipaa025>.

[777] Article 25 of the GDPR states that "taking into account the state of the art, the cost of implementation and the nature, scope, context and purposes of processing as well as the risks of varying likelihood and severity for rights and freedoms of natural persons posed by the processing, the controller shall, both at the time of the determination of the means for processing and at the time of the processing itself, implement appropriate technical and organisational measures, such as pseudonymisation, which are designed to implement data-protection principles, such as data minimisation, in an effective manner and to integrate the necessary safeguards into the processing in order to meet the requirements of this Regulation and protect the rights of data subjects".

[778] Article 46, §2º of the LGPD states that technical and organizational security measures must be observed from the design phase of the product or service to its execution.

[779] Examples of these principles include the principles of purpose specification, adequacy, necessity, data quality, transparency, security, non-discrimination, and accountability. All those can be found in Art. 6 of the LGPD. Other data protection laws such as the GDPR have similar lists of principles.

Having said that, it is important to emphasize that the methodology for sandbox experimentation remains similar to that of traditional sectors. The steps mentioned in the previous chapter are still present in DPAs' sandbox programs. What changes is the regulatory approach, and, in turn, the objective entailed in the use of this regulatory tool.

When moving towards the use of sandboxes in the regulation of artificial intelligence, it is necessary to question which approach one intends to adopt–licensing regimes models or approaches based on risk managing and fundamental rights protection. This will be further discussed in the next chapter.

## 18.4 Sandboxes And The Regulation Of Artificial Intelligence – Legislative Debates In Brazil And In The European Union

To discuss the most appropriate approach for artificial intelligence sandboxes, it is relevant to verify how this tool has been discussed in legislative proposals for the regulation of this technology. Thus, two cases will be analysed – the European and the Brazilian proposals.

The EU proposal, known as the AI Act,[780] is that of risk-based regulation, so that it is not necessary for a private or public sector actor to obtain a specific license for the use of artificial intelligence systems, if it complies with the rules established by the norm. According to the AI system's level of risk, a distinct set of obligations must be observed.

Regarding sandboxes, the AI Act creates an obligation to Members States to develop regulatory sandboxes in accordance with article 53.[781] Furthermore, it empowers competent authorities[782] to establish a regulatory sandbox to "provide a controlled environment that facilitates the development, testing and validation of innovative AI

[780] European Commission. (2021). *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 Com/2021/206 Final.* Brussels: European Commission. Retrieved from: <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52021PC0206>.

[781] AI Act, Article 53: "Member States shall establish at least one AI regulatory sandbox at national level, which shall be operational at the latest on the day of the entry into application of this Regulation This sandbox can also be established jointly with one or several other Member States".

[782] "Competent authorities" are the regulatory authorities of the AI Act, which may be defined on a case-by-case basis by each Member State of the European Union.

systems for a limited time prior to their placing on the market", as stated in Article 53(1).

On the other side of the Atlantic, inspired by the AI Act, the Brazilian National Congress has been discussing an AI Bill.[783] Bill n. 2,338/2023 is the main proposal, resulting from an intense debate held in 2022, when a Committee of Jurists was constituted. After six months of research, which included a comparative study on the experiences of the countries of the Organization for Economic Co-operation and Development – OECD, in AI regulation and several public hearings with national and international experts, the replacement text was presented in December 2022. This text was converted, in May 2023, into Bill n. 2,338/2023.[784]

Similar to the European proposal, Bill n. 2,338/2023 adopts a risk-based approach, in which AI systems have different levels of obligations, according to the risk classification. Furthermore, the way of listing high-risk AI uses is similar between the two frameworks. However, while the EU AI Act chooses to generate exhaustive lists of use, PL 2338/2023 opts for non-exhaustive lists. Therefore, while the European debate opted to limit the application of the law, which can be expanded in restricted cases, the Brazilian debate opted to amplify the application of the law, enhancing interpretative expansion by regulatory bodies. The listed use cases are very similar on both proposals.

However, unlike the version proposed by the European Commission,[785] the Brazilian proposal establishes a rights protection regime. Thus, Bill n. 2,338/2023 proposes rights for individuals affected by AI

[783] Shimoda, C. A.; Moraes, T. (2023). Brazil's Path to responsible AI. *OECD.AI Wonk Blog*. Retrieved from: <https://oecd.ai/en/wonk/brazils-path-to-responsible-ai.

[784] Brazil. Senado Federal Do Brasil. (2023). *Projeto de Lei nº 2.338, de 2023, de autoria do Senador Rodrigo Pacheco*. Brasília: Senado Federal do Brasil. Retrieved from: <https://www25.senado.leg.br/web/atividade/materias/-/materia/157233.

[785] The first version of the AI Act, proposed in 2021 by the European Commission, did not provide rules for the rights of individuals affected by AI systems. The version proposed in June 2023 by the European Parliament (EUROPEAN UNION, 2023), presents some rights, such as the right to lodge a complaint with a national supervisory authority, the right to an effective judicial remedy against a national supervisory authority, and the right to explanation of individual decision-making. European Parliament. (2023). *Amendments adopted by the European Parliament on 14 June 2023 on the proposal for a regulation of the European Parliament and of the Council on laying down harmonized rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts (COM(2021)0206 – C9-0146/2021 – 2021/0106(COD)*. Brussels: European Parliament. Retrieved from: <https://www.europarl.europa.eu/doceo/document/TA-9-2023-0236_EN.pdf.

systems, such as the right to preliminary information for individual interactions with AI systems; the right to an explanation of the AI system's decision, recommendation or prediction; the right to non-discrimination and correction of discriminatory effects, whether direct, indirect, illegal or abusive; and the right to privacy and protection of personal data, under applicable law. All these rights are detailed in specific provisions. Meanwhile, in the European text, these topics seems to be spread out throughout the provisions, embedded on other obligations.

This differentiates the Brazilian proposal, which is much closer to a regulatory framework of personal data protection, from what is seen in the European Commission's proposal which, while also adopting a risk-based approach, did not prescribe specific rights for individuals affected by AI systems. In any case, both proposals establish a set of obligations which indicate the importance of developing and using ethical and responsible AI systems that are transparent, allow human review and are safe enough to avoid incidents.

Curiously, in the chapter devoted to regulatory sandboxes, the AI Act seemed to come closer to data protection regimes. As several cases of AI systems classified as high-risk by the AI Act involve the processing of personal data, the European legislator was concerned that AI sandboxes comply with data protection legislation. Thus, the proposal presents, in article 54, conditions for the processing of personal data in AI sandboxes. In addition to limiting the scope of further processing to specific circumstances in the public interest (crime prevention, public safety, public health and environmental protection), the provision requires the implementation of various safeguards, such as the existence of effective monitoring mechanisms to identify whether high risks to the fundamental rights of data subjects may arise during testing, in addition to the isolation of personal data processing environments during experimentation.

Furthermore, as provided in Article 53(2), DPAs must be involved in the sandboxes, regardless of whether they are designated as an AI competent authority, when the innovation being tested involves the processing of personal data. This proposal makes DPAs the guardians of AI regulatory sandboxes in the EU.

Bill n. 2,338/2023, on the other hand, includes provisions[786] on the development of regulatory sandboxes by any regulatory authority, if authorized by a central AI supervisory authority, to be defined by the Executive Branch. The sandbox should provide information about the benefits that its participants will bring to consumers and society, as well as exit plans. The central AI supervisory authority can stop a program conducted by another regulator if it detects risks or damage to fundamental rights, including the protection of personal data.

Nevertheless, the Bill does not propose that the Brazilian DPA, ANPD, should be involved in other authorities' AI sandbox programs when those include high-risk systems that handle personal data. This gap could bring regulatory risks of legal certainty regarding compliance to the LGPD. In addition, the goals of the sandbox programs prescribed in the Bill focus only on fostering innovation, without establishing that it must pay attention to ethical and responsible values. While one may argue that the Bill provides as part of its fundamentals (art. 2), the rights of privacy and data protection and the respect for human rights, the specific provisions regarding sandboxes do not bring parameters on how this could be implemented. These are aspects that need to be reflected throughout the legislative debate to ensure that the future Brazilian AI regulation law achieves the objective of fostering innovation while protecting the rights of individuals affected by AI systems.

Anyway, considering the objectives and regulatory approaches of the proposals for the regulation of AI, both in Brazil and in the European Union, there seems to be greater alignment with the sandbox programs implemented by data protection authorities, which focus on fostering responsible innovations.

In this sense, it is worth noting that, on several occasions, AI systems were tested by the DPA sandboxes. In some cases, such as with ICO and PDPC, AI appeared incidentally, as part of the technologies being tested. However, Datatilsynet and SIC programs have focused specifically on AI-driven technologies. Both highlighted the importance of developing reliable AI systems which observed the implementation of ethical values such as those fostered by privacy by design.

[786] Arts. 39 to 42 of the Bill.

At the end of the program, reports are produced by the DPAs and sometimes by the participants, to share the observed good practices with non-participating entities that are developing similar innovations, and to spread the benefits of the innovations with society. For example, ICO (2023)[787] and Datatilsynet (2023)[788] provide reports from companies that participated in previous editions of the sandbox on their website.

Although the debate on the objective of AI regulatory sandboxes is far from being closed, it is worth bringing a last reflection in this article that will be important for any regulator that uses this instrument – its alignment with the institution's regulatory strategy.

## 18.5 Integrating Ai Sandboxes Into Regulatory Strategies – Synergies With (Really) Responsive Regulation

Before advancing, it is important to emphasize that sandboxes were not originally thought to be fitted as instruments of responsive regulation theory. As will be seen, the theory focuses on the supervisory role of the regulator, while the sandboxes operate in an earlier moment of experimentation and observation of the regulatory environment. That said, some relationship seems to exist, since experiments in the sandbox can induce behaviours in the regulated in order to direct them towards the desired regulatory compliance, as was presented in the cases of DPAs.

Given the transversal nature of artificial intelligence technology, it may be inevitable that sandbox programs implemented by regulatory authorities directly or indirectly involve the use of this technology. For example, in 2021, the global fintech market was responsible for moving 9,45 billion dollars in AI investments.[789] AI also fits into all the sectors mentioned in this study, and many more.

[787] ICO. Regulatory Sandbox – Previous Participants. London: ICO. 2023. Retrieved from: <https://ico.org.uk/for-organisations/regulatory-sandbox/previous-participants/>.

[788] Datatilsynet. Reports. Oslo: Datatilsynet. 2023. Retrieved from: <https://www.datatilsynet.no/en/regulations-and-tools/sandbox-for-artificial-intelligence/reports/>.

[789] Grand View Research. (2022). *Artificial Intelligence In Fintech Market Size, Share & Trends Analysis. Report By Component (Solutions, Services), Deployment (Cloud, On-premise), By Application (Fraud Detection, Virtual Assistants), And Segment Forecasts, 2022 – 2030*. California: Grand View Research. Retrieved from: <https://www.grandviewresearch.com/industry-analysis/artificial-intelligence-in-fintech-market-report>.

Therefore, the Brazilian legislator's proposal seems reasonable when allowing AI sandbox programs to be developed in the regulatory context of each authority, according to the regulated sector. At the same time, general guidelines would be issued by a central competent authority, so that all developed programs respect the protection of rights provided by the future AI legislation.[790] This authority would also be responsible for authorizing regulatory sandboxes on AI in Brazil (art. 38). Thus, this author argues that these guidelines should focus on ensuring that the various regulatory sandbox programs involving the use of AI foster the development of ethical and responsible innovations, regardless of any other goals which each regulator intends to achieve. Therefore, regulators must reflect on how sandboxes should be integrated into their regulatory strategies.

Far from exhausting the debate, this paper only brings a brief analysis of the apparent synergies of regulatory sandboxes with strategies based on the theory of responsive regulation. This theory has been widely adopted by national regulatory authorities in the face of the challenges brought by the growing dynamism and complexity of several sectors. For example, ANATEL has been adopting strategies based on this theory for some years, based on a study on legal theories of regulation supported by incentives.[791] In the financial sector, academic studies have reflected on correlations of strategies taken by the Brazilian National Financial System (SFN) with responsive regulation, such as its 2018 cybersecurity policy,[792] or the potential use of regulatory sandboxes by SFN entities to experiment with blockchain and anticipate risks related to money laundering.[793] In

[790] It is important to highlight that, albeit both the Brazilian and the EU bills propose a national central authority for AI governance, in the European case these authorities are parte of a wider set of governance bodies, which include a general council, reporting bodies and others.

[791] Aranha, Lopes, 2019 (op. cit.).

[792] In his paper, Goettenauer assesses whether there are elements of responsive regulation in the cybersecurity regulation of the SFN, presented in Resolution n. 4658, of April 26, 2018. The study concludes that it seems to be a new position by BACEN regarding the structuring of financial business in a digital environment, in the sense of responsive regulation, although it had not be yet fully adapted to the business context. Goettenauer, C. (2019). Regulação Responsiva e a Política de Segurança Cibernética do Sistema Financeiro Nacional. *Journal of Law and Regulation*, 5(1), 131–146. Retrieved from: <https://periodicos.unb.br/index.php/rdsr/article/view/20944>.

[793] Chagas, C. (2022) Lavagem de Capitais e a Blockchain: métodos alternativos de regulação. *Novas fronteiras do Sistema Financeiro Nacional.* Belo Horizonte: Ed. Expert, 413-442. Retrieved from: <https://pos.direito.ufmg.br/downloads/Novas-fronteiras-do-sistema-financeiro-nacional.pdf>.

addition, ANPD was inspired by this theory to develop its supervision and sanctioning administrative proceedings.[794]

As mentioned in this paper's introduction, Collingridge's dilemma presents the regulator with a complex challenge of defining when to regulate. Along with this challenge, Baptista and Keller (2016)[795] find that the decision on when to regulate is fatally connected to that on how to regulate. To this end, the regulator must decide not only the regulatory tools it will use, but also the strategies it will adopt in the use of these instruments.

In this sense, Wansley (2016)[796] argues that, for the regulation of technological innovations, the adoption of an experimental regulatory model is appropriate, as it allows the regulator to test the use of these technologies until obtaining satisfactory knowledge about the best regulatory measure to be adopted. According to the author, the experimentalist model aims to "maximize the potential for regulatory learning while preserving regulatory options". In addition, this approach mitigates the risk of "entrenchment" by political groups or social norms, since the longer an innovation is on the market, the greater the leverage power of lobbying groups and the stronger public opinion for its maintenance, bringing challenges to the regulator.

As they are considerably new, regulatory sandboxes seem not to have been mentioned in studies of regulatory theories yet. However, they seem to be a regulatory tool compatible with modern regulatory theories, such as Ayres and Braithwaite's (1992)[797] responsive regulation, and Baldwin and Black's (2008)[798] truly responsive regulation.

[794] As provided for in Resolution CD/ANPD n. 1, of October 28, 2021. Autoridade Nacional de Proteção de Dados – ANPD. (2021). *Resolução CD/ANPD nº 1, de 28 de outubro de 2021.* Brasília: ANPD. Retrieved from: <https://www.in.gov.br/en/web/dou/-/resolucao-cd/anpd-n-1-de-28-de-outubro-de-2021-355817513>.

[795] Baptista, P.; Keller, C. I. (2016) Por que, quando e como regular as novas tecnologias? Os desafios trazidos pelas inovações disruptivas. *Revista de Direito Administrativo*, 273, 123-163.

[796] Wansley, M. (2016). Regulation of emerging risks, *Vanderbilt Law Review*, 69(401). Retrieved from: <http://ssrn.com/ abstract=2646316>.

[797] Ayres, I.; Braithwaite, J. (1992). *Responsive regulation: Transcending the deregulation debate*. Oxford Walton Street: Oxford University Press.

[798] Baldwin, R., & Black, J. (2008). Really Responsive Regulation. *The Modern Law Review*, 71(1), 59–94. doi:10.1111/j.1468-2230.2008.00681.x.

Ayres and Braithwaite developed the theory of responsive regulation to transcend the impasse between "rigid" regulation and deregulation.[799] In short, they intend to find the right balance between punishment and persuasion to make regulation effective. This balance is sought addressing regulatory approaches in two complementary pyramids: the pyramid of supports and the pyramid of sanctions. These pyramids work in parallel, considering a carrot and stick approach – when the regulator wants to encourage a certain behavior (e.g., compliance duties) it concentrates on the first pyramid; when punishment is necessary, the last is used. The approaches follow a gradual escalation from bottom to top, which means that regulators should, in principle, start from the base of the pyramids. Aranha and Lopes (2019),[800] based on Braithwaite, presented an example of such pyramids (see Figure 1).

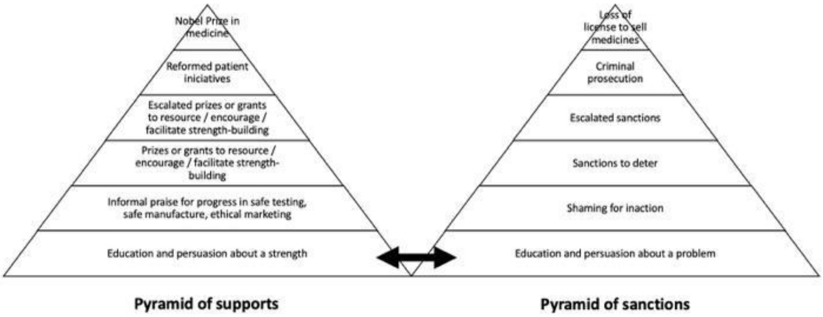

Nobel Prize in medicine
Reformed patient iniciatives
Escalated prizes or grants to resource / encourage / facilitate strength-building
Prizes or grants to resource / encourage / facilitate strength-building
Informal praise for progress in safe testing, safe manufacture, ethical marketing
Education and persuasion about a strength
Pyramid of supports
Loss of license to sell medicines
Criminal prosecution
Escalated sanctions
Sanctions to deter
Shaming for inaction
Education and persuasion about a problem
Pyramid of sanctions

Figure 1

Braithwaite's pyramids of support and sanction for the theory of responsive regulation (adapted from Aranha & Lopes, 2019, p. 232).

The support pyramid focuses on business compliance and continuous improvement, proposing several approaches that support the regulatees' training. In turn, the sanctions pyramid follows a more traditional command-and-control approach. Although regulatory sandboxes are not mentioned in the original proposal of responsive regulation theory, they seem to have some synergy with the bases of both pyramids, where education and persuasion are used to raise

[799] Pereira, J. R. (2022). *Openness Doesn't Hurt: Enforcing Qualified Machine-Learning Transparency For Data Protection Through Responsive Regulation*, Brasília: Universidade de Brasilia.

[800] Op. cit.

awareness of the strengths and problems of a given business model. The persuasive approach can be seem in the example of the DPA's sandboxes, in which, on several occasions, knowledge leveraging sessions were organized to share best practices on data protection and privacy by design principles. Ayres and Braithwaite (1992)[801] suggest that the regulator, as a starting point, should always have cooperation in mind.

According to Baldwin, Cave and Lodge (2012),[802] one of the challenges of responsive regulation is to ensure that there is clear communication between regulators and regulated, so that each one can understand the strategies adopted by the others. The existing interaction in sandbox environments is fruitful for this better communication, since the regulators get to know in advance the innovations that are being developed, and the regulated ones can understand the main points of concern of the regulators.

In an effort to improve the theory of responsive regulation, Baldwin and Black (2008)[803] present some criticisms in their paper named "Really Responsive Regulation". For example, they contest that step-by-step climbing is not always adequate, since, in certain cases, it will be necessary to start working at different points of the pyramid, or to advance more quickly. They also criticize the fact that the theory does not prepare the regulator to know in advance the behavior of the regulatee, which may not respond to pressure from the regulator, due to the market's culture. Another problem is that the theory assumes a binary relationship between regulator and regulated, when in fact it is necessary to consider a more complex ecosystem, in which there are several regulators acting concomitantly, including quasi-regulatory agents (such as civil society entities and market actors with dominant power who are capable of inducing behavior in other regulated countries).

Therefore, Baldwin and Black present a new version of the theory, named Really Responsive Regulation. Instead of pyramids, the authors propose a matrix analysis. On the horizontal axis, five elements

[801] Op. cit.

[802] Baldwin, R.; Cave, M.; Lodge, M. (2012). *Understanding regulation: theory, strategy, and practice*. Oxford: Oxford University Press, 2 ed.

[803] Op. cit.

must be considered: (i) behavioural attitudes of the regulatee; (ii) institutional regulatory environment; (iii) logical differences in regulatory strategies and tools; (iv) performance of the regulatory regime; (v) changes in the regulatory scenario. Really responsive regulation must be able to answer questions associated with these five elements.

To this end, they propose an analysis methodology that composes the vertical axis of the matrix. Thus, in each of the aforementioned elements, the regulator must be able to: (i) *detect* undesirable or non-compliant behavior; (ii) *respond* to behavior based on the development of rules and tools; (iii) *enforce* the tools based on outlined strategies; (iv) *analyse* the success or failure of the strategies and tools implemented; (v) *modify* the strategies and tools according to the observed result. In this way, the regulator will always be feeding back its regulatory strategy, to adapt to continuous changes in the regulatory environment.

For regulatory sandboxes to be useful tools for the regulator, it is necessary to reflect on how they fit into the regulatory system. On another paper, Black (2021)[804] proposes an analytical framework for these systems, consisting of six key elements that interact with each other constantly, to produce a dynamic system: (i) objectives, purposes and values; (ii) knowledge and understandings; (iii) tools and techniques; (iv) behaviours; (v) organizations, structures and processes; and (vi) trust and legitimacy.

Sandboxes fill the third element of Black's framework. It would be interesting for the regulator to reflect on how this tool interacts with others, as well as with other elements of the framework. Sandboxes can be useful to increase the regulator's knowledge on innovations being developed by the regulatee, as well as to induce regulatees behaviours in the development of responsible innovations.

Baldwin and Black's proposals seem to have a lot of synergy with benefits in implementing regulatory sandboxes. DPAs' experiences indicate that sandboxes can be excellent instruments for observing

[804] Black, J. (2021). Constitutionalising Regulatory Governance Systems, *LSE Law, Society and Economy Working Papers*. Retrieved from: <https://eprints.lse.ac.uk/113670/1/Black_constitutionalising_regulatory_governance_published.pdf>.

and inducing behavior in regulated parties, either by encouraging compliance with the principles inherent in data protection legislation during testing, or by replicating behavior in market players who did not participate through the sharing of lessons learned in public reports. Furthermore, as mentioned by the German government, sandboxes allow the regulator to better understand the regulated environment and to be able to develop future norms that are more aware of the reality of the market.

## 18.6 Conclusions – The Future Of Artificial Intelligence Sandboxes And Their Integration Into Regulatory Strategies

The reflections brought by this study are far from closed. However, they point to a convergence in the use of sandboxes as relevant tools for fostering innovation in regulatory ecosystems. Considering the strong correlation between data protection legislation and the artificial intelligence regulatory regimes that have been proposed in Brazil and in the European Union, it would be ideal if future AI regulators also aim to develop sandboxes that foster ethical and responsible innovations, and not only innovation for its own sake.

It is also crucial to keep in mind that regulatory sandboxes alone will not solve all regulatory problems. The regulator should prefer policy mixes incorporating combinations of institutional tools. This means that sandboxes must be considered part of a set of instruments that the regulator must use according to its incentives and restrictions' strategy. Therefore, it will never be a substitute for administrative sanctions or other instruments of persuasion.

Regulatory sandboxes which focus on inducing ethical behavior to the regulated may be able to find a balance between the diminishing of regulatory barriers to innovation and the prevention of social harm, making them excellent tools to foster the development of responsible innovations, such as trustworthy AI systems. The path proposed by DPAs points to the use of these tools to promote ethical and responsible innovation. It remains to be seen what kind of innovation AI regulators will encourage.

# 19 Building a repository of public algorithms: Case study of the dataset on automated decision-making systems in the Colombian public sector

**Juan David Gutiérrez, Profesor Asociado de la Escuela de Gobierno Alberto Lleras Camargo de la Universidad de los Andes;**

**Sarah Muñoz-Cadena, estudiante de la maestría en Economía de las Políticas Públicas de la Universidad del Rosario e investigadora en Policéntrico.**

## Abstract

This article documents how the team of scholars built the new repository of public algorithms in Colombia and describes how the data was collected, processed, and organized. The article also explains the main difficulties that the researchers encountered as well as the solutions that were implemented. Finally, the article reflects on the challenges of fostering algorithmic transparency in a Global South country and offers recommendations for replicating this project in other countries. The dataset comprises 113 automated decision-making systems (ADMs) of 51 Colombian public bodies, characterizes each system with 40 variables, and was built with over 300 different sources of publicly available information. The ADMs are used by public organizations in Colombia to perform a wide range of functions and to support different types of state activities, but almost half of them are concentrated in the justice, education, and environment sectors.

## 19.1 Introduction[805]

Colombia's repositories of public algorithms document less than 30 automated decision-making systems (ADMs) despite of the fact that 233 public bodies answered, in a government-led survey in 2021, that they used artificial intelligence (AI) and/or robotic process automation (RPA) systems.[806] The apparent under-registration of

[805] The results of the research presented in this paper are part of a broader project that studies the life cycle of ADMs projects in the Colombian public sector, as well as the implications of these systems for public management, democracy, and society. Therefore, although this paper is novel, the methodology of the dataset that was built and some of the descriptive statistics were included in an article recently published by the authors.

[806] Gutiérrez, J. D., & Muñoz-Cadena, S. (2023). Adopción de sistemas de decisión automatizada en

ADMs by the Colombian state led a team of scholars, financed by Universidad del Rosario, to build a new dataset that mapped the systems adopted by the public sector.

After one year and a half of work, in June 2023 the researchers published a new database of 113 ADMs from 51 Colombian public sector organizations. The database characterizes the systems based on 40 variables and was created with more than 300 sources of publicly available information (Gutiérrez et al., 2023).[807]

This article documents how the team of scholars built the new repository of public algorithms in Colombia and describes how the data was collected, processed, and organized. The article also explains the main difficulties that the researchers encountered as well as the solutions that were implemented. Finally, the article reflects on the challenges of fostering algorithmic transparency in a Global South country and offers recommendations for replicating this project in other countries.

This article is divided into four sections, including this introduction. The second section offers a brief account of the state-of-the-art on repositories of public algorithms published by national and subnational governments. The third section describes the methodology used to build the new Colombian dataset and describes our main findings. The last section discusses the conclusions and recommendations about building repositories of public algorithms.

## 19.2 State-of-the-art on repositories of public algorithms around the world

Public repositories of algorithms are "windows" and "channels" where individuals can find information "to understand how the system works, how its decisions were done ('explainability') and to contest its behaviours ('accountability')",[808] and to understand where the data comes from and what results the system may produce ('traceability').

el sector público: Cartografía de 113 sistemas en Colombia. *Working Papers GIGAPP*; Gutiérrez, J. D., Muñoz-Cadena, S., & Castellanos-Sánchez, M. (2023). *Sistemas de decisión automatizada en el sector público colombiano [Dataset]* (Versión V1) [dataset]. Universidad del Rosario.<https://doi.org/10.34848/YN1CRT>.

[807] The database can be downloaded at the following link: <https://doi.org/10.34848/YN1CRT>.

[808] Haataja, M., van de Fliert, L., & Rautio, P. (2020). *Public AI Registers.Realising AI transparency and civic participation in government use of AI* [White paper]. <https://algoritmeregister.amsterdam.nl/wp-content/uploads/White-Paper.pdf>., p. 5.

Some supranational, national, and subnational governments around the world have created repositories that provide information on ADMs adopted by the public sector. Furthermore, a few civil society organizations and universities have also stepped up to publish datasets and to contribute with algorithmic transparency.[809] The following table summarizes sixteen repositories that are publicly available online, including the new dataset documented in this article:

**Table 1 – Repositories of public algorithms**

| Name of the repository | Geographical Scope | Organization that published it | Nature of the organization |
|---|---|---|---|
| **America** | | | |
| Algoritmos públicos | Chile | GobLab – Universidad Adolfo Ibáñez | University |
| Tablero de seguimiento marco ético | Colombia | National government | Public |
| Sistemas de decisión automatizada en el sector público colombiano | Colombia | Universidad del Rosario | University |
| Algorithmic tools. | New York, United States | Subnational government | Public |
| Artificial Intelligence and Algorithms | Ontario, Canada | Regional government | Public |
| Algorithm tips | United States | Northwestern University | University |
| Proyectos de aplicación de Inteligencia Artificial | Uruguay | National government | Public |
| **Europe** | | | |
| City of Amsterdam Algorithm Register. | Amsterdam, The Netherlands | Subnational government | Public |
| Inventaire des algorithmes utilisés par la Ville d'Antibes. | Antibes Juan-les-Pins, France | Subnational government | Public |
| Publication des algorithmes et des codes sources. | France | National government | Public |
| Artificial intelligence systems of Helsinki. | Helsinki, Finland | Subnational government | Public |
| Consultation des Algorithmes publics de Nantes Métropole. | Nantes, France | Subnational government | Public |

[809] A detailed list of these public algorithm repositories can be found at the following link: <https://forogpp.com/inteligencia-artificial-y-sector-publico/repositorios-y-registros-de-algoritmos/>.

| Name of the repository | Geographical Scope | Organization that published it | Nature of the organization |
|---|---|---|---|
| EU Selected AI cases in the public sector -AI-WATCH: EU Artificial Intelligence Observatory. | European Union | Supranational government | Public |
| AI-X: AI Public Services Explorer. | European Union | Supranational government | Public |
| Algorithmic Transparency Reports. | United Kingdom | National government | Public |
| **Global** | | | |
| Observatory of Algorithms with Social Impact – OASI. | Worldwide | Eticas Foundation | Civil society organization |

Source: Authors' own elaboration based on Gutiérrez (2023d).

Most of the repositories listed in Table 1 are part of broader algorithmic transparency initiatives, whereby governments aim at improving the accessibility and explainability of information related to the use of algorithms to (semi) automatize their decision-making processes. In the case of the repositories created by the Colombian government, the initiatives were related to the implementation of Open Government policies –more specifically open data projects– and the implementation of an AI Ethical Framework published by the national government.

There are also cases of public-private partnerships, such as the repository created by the GobLab of the Adolfo Ibáñez University (UAI) and the Council for Transparency of Chile (CPLT), an autonomous public organization that oversees the compliance of the Law on Transparency of the Civil Service and Access to Information of the State Administration. The repository was created through the joint work of GobLab and CPLT, it was first published in an online platform in November 2021, and currently maps over 90 ADMs in the Chilean public sector.[810]

[810] Garrido, R., Lapostol, J. P., & Hermosilla, M. P. (2021). *Transparencia algorítmica en el sector público*. GOB LAB UAI. Consejo para la Transparencia. <https://goblab.uai.cl/wp-content/uploads/2021/10/ESTUDIO-TRANSPARENCIA-ALGORITMICA-EN-EL-SECTOR-PUBLICO-GOBLAB-CPLT-final.pdf>; GobLab UAI. (2022). *Repositorio de algoritmos públicos de Chile. Primer informe de estado de uso de algoritmos en el sector público*. Universidad Adolfo Ibáñez (UAI). <https://goblab.uai.cl/wp-content/uploads/2022/02/Primer-Informe-Repositorio-Algoritmos-Publicos-en-Chile.pdf>; GobLab UAI. (2023). *Repositorio Algoritmos Públicos. Informe Anual 2023*. Universidad Adolfo Ibáñez (UAI).

The research that produced the new dataset for the Colombian public sector, documented in this article, also aimed to contribute with algorithmic transparency. However, it was not a process driven by the state, nor public-private partnerships, but an academic project that was financed Colombian private university and that aimed at generating knowledge on how the Colombian public sector uses ADMs. However, the design of our project and its implementation was inspired by the GobLab's repository in Chile, as we will explain in the following section.

## 19.3 How we built our dataset of ADMs in the Colombian public sector

### 19.3.1 Identifying the data gap

As we mentioned before, the Colombian national government has published repositories of public algorithms, but when we explored them in the second semester of 2021, we noticed that the number of systems that were registered seemed to be very low. The main clue about this under-recording was found in the answers of the public entities that filled out the "2021 Management Progress Report Form" (Formulario Único de Reporte de Avances de Gestión – FURAG).[811] In 2021, the FURAG was completed by 2,939 public entities and of these, a total of 233 (8%) answered that they used AI and/or robotic process automation (RPA) systems. More precisely, 172 (74%) reported using AI, 116 (50%) using RPA, and 55 (24%) using both types of technologies.[812]

However, it was likely that some of the public entities that responded affirmatively the FURAG survey about the use of IA and/or RPA systems did not actually use these types of systems. The terms IA and RPA are not commonly used by public officials and the Department of Public Function (DAFP) did not include these definitions in its most recent FURAG glossary,[813] hence number of entities that used ADMs would probably be lower.

[811] Information on the type of questions included in the FURAG and the collection methodology is available on the following DAFP platform: <https://www.funcionpublica.gov.co/web/mipg/medicion_desempeno.

[812] The FURAG response data can be consulted on the following platform: <https://www.datos.gov.co/Funci-n-p-blica/FURAG/daed-z4fw/data.

[813] The June 2020 glossary (version 5) of the FURAG is available at the following URL: <https://

Our new database confirmed the under-registration of algorithms in the government-based repositories, as we will explain in the following pages.

## 19.4 Data Collection

Following the advice of the project directors of the GobLab's who generously shared with us their experience building Chile's repository of algorithms, we searched for a governmental partner who was interested in jointly working on a new Colombian dataset. Between February and June 2022, we held several meetings with a national government agency to co-design the project. In the meantime, in May 2022, the team of researchers at Universidad del Rosario started to collect data about ADMs adopted by the Colombian state based on publicly available information.

Between July and August 2022, we agreed on a set of questions that the agency would send to the public bodies that had answered affirmatively the FURAG 2021 survey and designed the online forms that we would use to collect the data. In September 2022, the agency sent the communications to over 200 governmental organizations and until November 2022 we received information about 203 systems adopted by over 80 national and subnational governments. Hence, some governmental bodies had sent information about two or more systems and around a third of the organizations responded the request for information (rate of response). We found that around 60 of the systems that had been informed were not AI nor RPA (for example, some governmental bodies reported information systems, accounting software and antivirus software). Additionally, there was not publicly available information about approximately 100 systems that were reported by the governmental organizations.

In the October and November 2022, the researchers attempted to coordinate with the agency to send a second round of communications to government bodies that had not answered the initial request of information. Unfortunately, this second round of communications was never sent because the priorities of the governmental body

www.funcionpublica.gov.co/documents/28587410/36200637/Glosario_mipg.pdf/9ff42c08-61a9-e0fa-76b1-1f662c0b2202?t=1593207412671>.

changed in the last quarter of the year. Here it is important to mention that in August 2022 a new national government took office in Colombia. In the last days of November 2022, we were informed by the national government agency that they would suspend sending new communications related to our project given that the government had to concentrate in the work on the new national development plan 2022-2026 and that they would be available to resume the work at some point of 2023, but this proposal did not materialize.

Hence, we decided not to use that data to construct the repository despite of the valuable information jointly retrieved by the researchers and the national governmental agency. Part of the data that we had collected was not publicly available and we did not want to publish it without the explicit consent of the national government agency.

Fortunately, the researchers had already started a Plan B: building the dataset with publicly available information. In the first semester of 2023 we accelerated the search for pertinent information and started populating our dataset. By June 2023, we had collected data from more than 300 primary and secondary sources that were publicly available. To curate the database the three researchers that built it met on a weekly basis to discuss each of the variables used to describe the systems that we introduced in the database.

The main type of source we consulted was information published by public entities through public data repositories, annual management reports, institutional press releases, official websites, and post on their official social networks' accounts. In total, the database used 210 institutional sources (68%) to document the ADMs that were identified. The database also used secondary sources: 45 press articles (15%), 24 academic publications (8%), 11 documents from multilateral organizations (4%), nine documents from private companies (3%) and eight publications from civil society (2%).

The new database has information on 113 ADMs of national and subnational public entities that are part of the executive branch, the judiciary, and other autonomous agencies. This includes systems adopted by ministries, superintendencies, mayors' offices, judicial bodies, and state-owned public utilities, among others. We found that 97 systems are in operation (86%), 14 are still in a pilot a phase (12%), one is suspended, and one was discontinued.

The database characterizes each system with regards to 40 variables which can grouped into five categories: (i) basic information on the ADS, including the system' name or project, data on the public entity that implements it, main objectives, status of the system, among others; (ii) type of data that the system requires, including, but no limited to, if it uses personal data; (iii) information on the executor and financier(s) of the project, in addition to the amounts and where the resources come from; (iv) classification of the ADS according to the governance function of the public entity that adopts them, according to the sector to which it contributes, according to the type of functions of the tool, according to the stage or stages of the public policy cycle to which it contributes, and potential contributions to the Sustainable Development Goals (SDGs); and, (v) information on the primary and secondary sources used to characterize each system.

As we mentioned above, one important limitation of the new database is that it only documents systems for which there is public information. Due to the unfinished research carried out jointly with a national government organization we know that there are more ADMs that could be mapped. Moreover, it is very likely that we will never obtain information of some systems used to perform national security and/or defense functions given that their existence is confidential. Finally, the database does not include systems that can be accessed by any user, public or private, through the Internet. For example, we do not include in the database large-scale language models that can be used through chatbots, such as ChatGPT, and that have been used by Colombian judges and magistrates to take court decisions.[814]

[814] Gutiérrez, J. D. (2023a). ¿Están los jueces en capacidad de usar modelos de lenguaje a gran escala (LLMs)? *Revista EXCEJLENCIA*, *7*, 10-15; Gutiérrez, J. D. (2023b). *Hablemos sobre el uso de ChatGPT para redactar decisiones judiciales*. La Silla Vacía. <https://juangutierrez.co/publicaciones/otras-publicaciones/#:~:text=Hablemos%20sobre%20el%20uso%20de%20ChatGPT%20para%20redactar%20decisiones%20judiciales>; Gutiérrez, J. D. (2023c, febrero 23). ChatGPT in Colombian Courts: Why we need to have a conversation about the digital literacy of the judiciary. *VerfBlog*. <https://verfassungsblog.de/colombian-chatgpt>/; Gutiérrez, J. D. (2023d, julio 6). Repositorios y registros públicos de algoritmos. *Foro Administración, Gestión y Política Pública*. <https://forogpp.com/inteligencia-artificial/repositorios-y-registros-de-algoritmos/>.

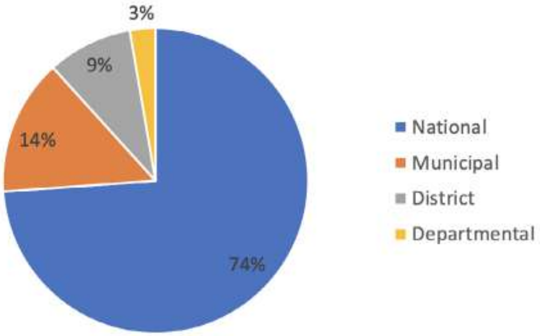
## 19.5 Main findings

Although this article focuses on the process of creating the dataset of ADMs adopted by the Colombian public sector, it is worth highlighting the most important findings that may be derived from it. The statistics presented below include 111 systems, because we excluded the system that is suspended and the discontinued one.

The public entities mapped in this research that are using ADMs in Colombia are, for the most part, from the executive branch (93%) and, to a lesser extent, from the judicial branch (4%) or belong to control entities (3%). We did not find any system adopted by the legislative branch.

The database records that 51 different public entities adopted ADMs. Of the 111 systems (see Figure 1), 82 (74%) were piloted or implemented by national entities and 29 (26%) by territorial entities (municipal, district or departmental).[815] In the case of local entities, these are distributed among initiatives led by municipal (14%), district (9%) and departmental (3%) entities.

**Figure 1. Percentage of systems adopted by national or territorial entities.**



Source: Authors' own elaboration.

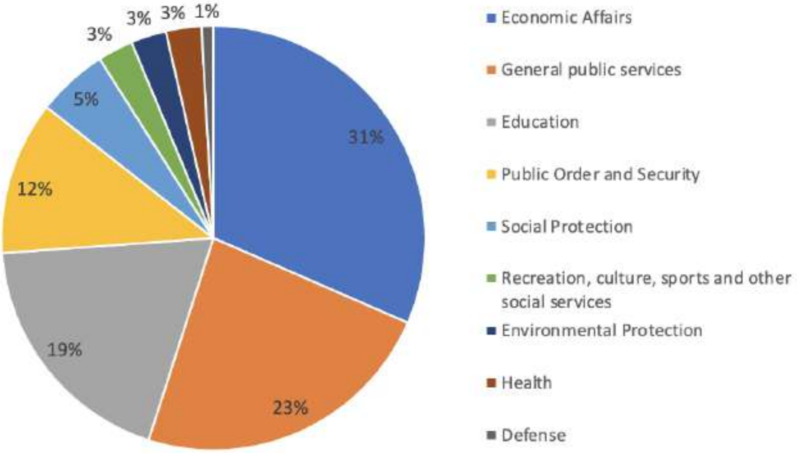
Applying the Classification of the Functions of Government (COFOG),[816] we found that 74% of the mapped systems were adopted by public

[815] In the case of ADMs used by different Secretariats of the Mayor's Office of Bogota, they are considered as a single entity in these statistics.

[816] "The Classification of the Functions of Government (COFOG) is a detailed classification of the functions and socioeconomic objectives pursued by general government units through different types of expenditure. It makes it possible to identify the expenditure made by the government in

entities that perform three categories of functions: "economic affairs" (32%), "general public services" (23%) and "education" (19%) (see Figure 2).

**Figure 2. ADS classification based on the first level of COFOG.**



Source: Authors' own elaboration.

This finding for the Colombian case contrasts with that reported in the GobLab UAI Public Algorithms Repository: in the Chilean case, of the 75 algorithms that are registered,[817] 61% are related to three sectors: "health" (25%), "economic affairs" (24%) and "public order and security" (12%). On the other hand, in a mapping of AI systems used by public entities in the European Union, in which COFOG is also applied, it was found that the three categories with the most mapped algorithms were: "general public services", "economic affairs", and "public order and safety".[818]

Returning to our database, COFOG allows us to detail the type of function performed by the public entities that adopted ADMs

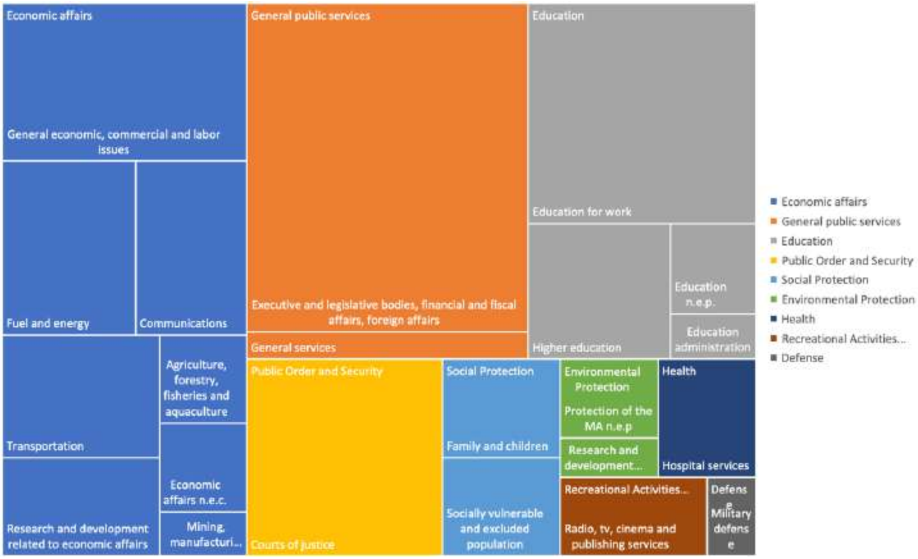
accordance with the purposes or public functions, showing the nature of the services provided by the institutions on behalf of the state", DANE. (2020). *Clasificación de las funciones del gobierno (COFOG)*. <https://www.dane.gov.co/files/sen/nomenclatura/cofog/COFOG-AC.pdf>.

[817] GobLab UAI. (2023). *Repositorio Algoritmos Públicos. Informe Anual 2023*. Universidad Adolfo Ibáñez (UAI).

[818] Misuraca, G., Van Noordt, C., & Boukli, A. (2020, septiembre 23). *The use of AI in public services: Results from a preliminary mapping across the EU*. Proceedings of the 13th International Conference on Theory and Practice of Electronic Governance (ICEGOV 2020), Athens, Greece. <https://doi.org/10.1145/3428502.3428513>, p. 95.

from a second level of categorization, which allows us to classify the entities based on more detailed government functions, as illustrated in Figure 3.

**Figure 3. Classification of the functions of the public bodies that adopted ADS.**

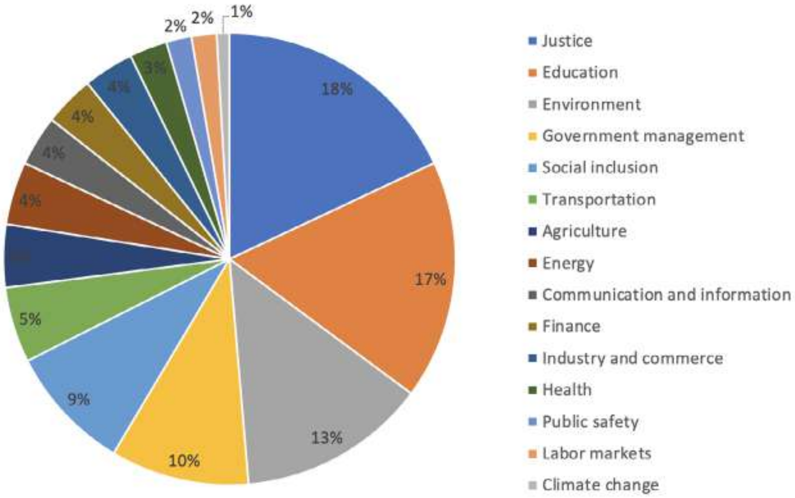


Source: Authors' own elaboration.

Although COFOG is useful because it is a standardized classification used by different countries to report on the functions pursued by government organizational units, which facilitates comparison between jurisdictions, its main limitation for this research is that it does not accurately report the type of government activity to which each ADMs contributes. This is because public entities may perform functions that could be placed in more than one COFOG category. For example, a public entity located in the "economic affairs" category, such as the Superintendence of Industry and Commerce, performs functions that are judiciary in nature.

Therefore, to complement the characterization of the ADMs, we use the categories used by the Inter-American Development Bank (IADB) in the framework of the fAIr LAC initiative. Specifically, the fAIr LAC Observatory developed a classification of 18 sectors to which AI initiatives in Latin America can contribute.[819] The ADMs are used by public organizations in Colombia to perform a wide range of functions and to support different types of state activities, but almost half of them are concentrated in the justice, education, and environment sectors.

**Figure 4. Classification of the ADMs by type of sector according to the classification used in the IDB's fAIR LAC Observatory.**



Source: Authors' own elaboration.

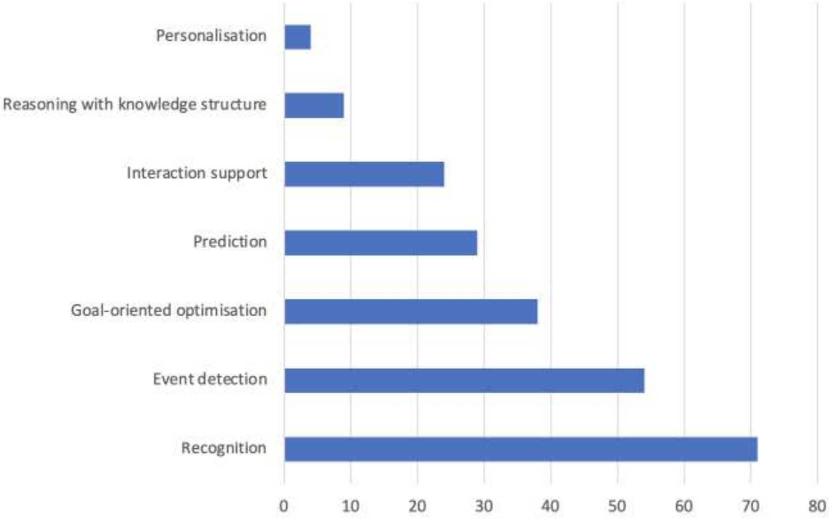
To classify ADMs according to the type of function performed by each tool, we applied the classification of the Organization for Economic Co-operation and Development (OECD), which identifies seven classes according to the type of output generated by the AI.[820] It is pertinent to mention that usually systems can perform more than one function. We found that 71 of the 111 ADMs (64%) perform recognition functions, 54 (49%) detect events, 38 (34%) focus on goal-focused optimization, 29 (26%) do event prediction, 24 (22%) provide interaction support (in particular, chatbots) nine

[819] The fAIr LAC Observatory is available at the following URL <https://fairlac.iadb.org/observatorio>. Although we use the fAIr LAC Observatory categories, we do not have any systems registered in four categories: aquaculture, gender or diversity, perspective and personal data protection. In the meantime, it should be clarified that we included an additional category, which was not included in the IADB classification: transportation.

[820] OECD. (2022). *OECD Framework for the Classification of AI systems*. OECD. <https://doi.org/10.1787/cb6d9eca-en>.

(8%) perform knowledge-structured reasoning and four (4%) seek to personalize (see Figure 5).

**Figure 5. Classification of ADMs in Colombia by type of functionalities.**



Source: Authors' own elaboration.

## 19.6 Reflections on building repositories of public algorithms

This article described how a group of researchers, financed by Universidad del Rosario, built a new database that maps the adoption of ADMs by the Colombian public sector. The project was initially devised to be undertaken jointly with a governmental body, but political and administrative processes changed the priorities of the public organization and led to a suspension of the joint project in late 2022.

Despite of this setback, the team of academics reconfigured the project and published in June 2023 a database that identifies 113 ADMs in the Colombian public sector. The systems are characterized with 40 variables and that is based on more than 300 sources of primary and secondary information.

We would like to close this article by sharing our reflections about the process of building the dataset that may be pertinent for other organizations around the world that may be interested in creating new repositories of public algorithms.

One of the first steps to comply with the principle of algorithmic transparency is informing citizens about the ADMs that the state uses, which public entities use them, what they use them for, and how the systems operate. In this sense, a public repository of algorithms becomes a first "window" for citizens to get informed. Many governments of the World, at different levels, are using ADMs but very few of them have proactively informed about their use through repositories of public algorithms.

In this respect, we consider that there is an important opportunity to improve algorithmic transparency through joint work of coalitions among the state, civil society organizations and academia, that can create accessible, detailed, and sustainable public repositories that are available online. One of the greatest challenges we faced in the process of building the database is the availability of the information. The challenge is not just to find that a system exists and that is being used, but also to know how the system was built (e. g. the data used to train if its AI) and how it operates. For example, we found that there was scant information about the costs of the system, how it was financed and, if the system is already in use, whether there is any reporting about the results obtained through its implementation.[821] Not being able to access such information makes impossible for third parties to assess the performance of the system.

In these information search processes, it is important not to limit the search only to the entities' annual management reports or official web pages; it is also important to search the official social media accounts of the public entities and the websites of the private companies that developed the algorithms. In fact, we found that when algorithms are developed through public-private partnerships, the private organizations tend to provide more extensive information about the system.

[821] The Colombian State has the Electronic System for Public Procurement (SECOP), whereby law the contracts signed by public entities should be registered. In our search for information on the costs of the systems, we used in the SECOP search engine the names of the systems and keywords such as "chatbot", "automated decision system", "robotization of processes", "algorithm", "machine learning", but we only found information for seven of the 111 systems in operation or piloting.

Public repositories require continuous work to ensure that the information contained in them is up to date, since for various reasons the systems registered there may be discontinued, or new systems may be implemented. In this sense, it would be worth that states consider issuing basic mandatory rules for organizations that use ADMs in the public sector. This is what the CPLT of Chile is currently working with the support of GobLab, a general instruction that obliges government organizations to disclose key information about the ADMs they operate.

Finally, we would like to finalize sharing good news: we are currently working with the national government agency to resume the project and work on a new database in 2024.

# 20 Esfuerzos Internacionales destinados a promover la transparencia y/o la rendición de cuentas de la IA

**Jesús Javier Sánchez García, National Institute of Transparency, Access to Information and Personal Data Protection (INAI Mexico);**

**Nadia Elsa Gervacio Rivera, INAI Mexico;**

**Jonathan Mendoza Iserte, Secretary of Personal Data Protection, INAI Mexico.**

*“Si se espera que una máquina sea infalible, tampoco puede ser inteligente”.*

*Alan Turing*

## Resumen

La Inteligencia Artificial (AI) supone beneficios, pero también plantea retos para su regulación, por lo que es fundamental conocer cuáles han sido los esfuerzos internacionales que a la fecha han buscado regularla, destacando la importancia de promover los principios sobre transparencia y rendición de cuentas en la creación e implementación de tecnologías que utilizan sistemas de inteligencia artificial. Además, se planteará el panorama de la gobernanza responsable de la IA para generar confianza en los usuarios, a través de la implementación de sus principios tanto a nivel nacional como internacional, con el fin de maximizar los beneficios para la sociedad y minimizar los riesgos potenciales de su uso, mediante la colaboración, entre países y partes interesadas para la promoción de una IA confiable y ética.

El planteamiento final de este documento se centra en la presentación de una propuesta latinoamericana basada en la jurisprudencia internacional existente encaminada a la creación de un mecanismo ex profeso que coadyuve en asuntos relacionados con la inteligencia artificial en esta región, mediante la cooperación y el establecimiento de alianzas estratégicas con organizaciones internacionales como la Organización de los Estados Americanos (OEA), a través del Comité Jurídico Interamericano, y con el apoyo y participación de otros bloques económicos a nivel global que han mostrado su interés en el tema.

## 20.1 Introducción

Los esfuerzos internacionales destinados a la regulación de la inteligencia artificial (IA) están cobrando cada vez más importancia en un mundo interconectado y dependiente de la tecnología. Diversos organismos y coaliciones internacionales han surgido y se han unido con el objetivo de establecer directrices y normativas que guíen el desarrollo y la aplicación de la IA de manera responsable para mejorar sus estándares y marcos regulatorios que promuevan la transparencia, la imparcialidad, la privacidad y la inclusión en sus procesos.

A medida que se acelera el ritmo de la evolución tecnológica hemos visto cómo se ha fortalecido a las sociedades y a las economías mejorando las condiciones de vida y aumentando la productividad, es por esto que debemos resaltar la importancia de abordar los desafíos comunes de la gobernanza tecnológica global con el fin de identificar posibles brechas sin detener el curso de la innovación.

Por lo antes expuesto, este trabajo tiene un enfoque basado en la teoría del multilateralismo, tomando en cuenta la dinámica internacional que ha contribuido a la transformación vivida en años recientes, resultado de los avances que ha traído consigo la innovación tecnológica que ha impactado los ámbitos locales, regionales y el global afectando de manera positiva o negativa a nuestras sociedades.

Por definición entendemos que el multilateralismo es concebido como “una serie de arreglos transitorios entre más de dos Estados que, habiendo encontrado puntos de interés común, se proponen transformarlos en objetivos y acciones colectivas”[822] (Cox, 1996).

Aunado a la concepción de Estados a los que hace referencia esta teoría, con el tiempo se fueron sumando actores del escenario internacional que han jugado un papel clave en el desarrollo e implementación de la cooperación internacional y las prácticas multilaterales efectivas, como son los organismos internacionales y la sociedad civil organizada.

Lo anterior podemos verlo reflejado en el ímpetu y la proactividad mostrada por diversas organizaciones internacionales que han

[822] Cox, Robert W. *Multilateralism and World Order,* en R. W. Cox y T. S. Sudair (eds.): Approaches to World Order, Cambridge University Press, 1996.

promocionado guías, orientaciones y lineamientos específicos para el uso responsable de los sistemas de inteligencia artificial, siempre con estricto apego al respeto de los derechos humanos con un enfoque centrado en el ser humano. Ejemplo de lo anterior son los ejercicios que se han puesto en marcha como la Recomendación del Consejo sobre Inteligencia Artificial emitida por la Organización de Cooperación y Desarrollo Económico (OCDE); la Recomendación sobre la ética de la inteligencia artificial de la Organización de las Naciones Unidas para la Educación, la Ciencia y la Cultura (UNESCO), la Ley de Inteligencia Artificial de la Unión Europea y de los grupos de las potencias económicas mundiales que se han pronunciado al respecto a través del Comunicado de Líderes del G7 de Hiroshima y la Declaración Ministerial Reunión de Ministros de Tecnología y Digital del G7, así como el comunicado del grupo de compromiso del G20.

## 20.2 Esfuerzos internacionales destinados a la regulación de la IA

La colaboración internacional es esencial para la IA, ya que no conoce de fronteras y su impacto se extiende a múltiples sectores, desde el sector salud hasta cuestiones relacionadas con el medio ambiente. A tal efecto, para garantizar un desarrollo equitativo y ético de la IA, es crucial que los países y regiones intercambien conocimientos y buenas prácticas. Esto debe hacerse con apego al derecho internacional y promoviendo la cooperación entre países.

Los esfuerzos normativos a la fecha son insuficientes y no se cuenta con una regulación vigente en la materia encargada de vigilar y prevenir los riesgos latentes, a pesar de la propuesta existente de la ley de IA en Europa que se estima entrará en vigor hasta 2026[823] (CNN Business, 2023). Ninguna de las iniciativas recientes que provenienen de los organismos internacionales es vinculante, por lo que su alcance e impacto es flexible y voluntario a manera de recomendaciones o principios que no garantizar el éxito de la gobernanza de la IA.

[823] Ziady, Hanna. (2023). Europe is leading the race to regulate AI. Here's what you need to know. *CNN Business.* <https://edition.cnn.com/2023/06/15/tech/ai-act-europe-key-takeaways/index.html>.

Aun así, estos ejercicios conforman el primer conjunto de instrumentos normativos internacionales para afrontar los desafíos que representa la IA, a través de un marco de gobernanza sugerido que brinde soluciones centradas en la estabilidad global y el equilibrio tecnológico.

### 20.2.1 OCDE

El primer instrumento a nivel global fue "La Recomendación del Consejo sobre Inteligencia Artificial" adoptada en mayo del 2019 (OCDE, 2019).[824] La recomendación establece principios prácticos y flexibles para que sigan vigentes a través del tiempo y que complementan otros estándares como la privacidad, la gestión de riesgos de seguridad digital y la conducta empresarial responsable.

### 20.2.2 G20

Posteriormente, en junio de 2019 el G20 adoptó los Principios de IA centrados en el ser humano que se basan en los Principios de IA de la OCDE, a grandes rasgos señalan que teniendo en cuenta cómo la tecnología afecta a la sociedad se debe proporcionar un entorno propicio para el desarrollo centrado en el ser humano.

Este documento muestra que las tecnologías de IA pueden ayudar a promover el crecimiento económico inclusivo y brindar grandes beneficios a la sociedad; su uso responsable puede fortalecer los Objetivos de Desarrollo Sostenible (ODS),[825] (G20, 2019).

### 20.2.3 UNESCO

En noviembre de 2019 la UNESCO publicó la Recomendación sobre la ética de la Inteligencia Artificial dirigida a los Estados Miembros, en su calidad de actores de la IA y como autoridades responsables de la elaboración de marcos jurídicos y reguladores de los sistemas de IA. También proporciona orientación ética a todos los actores de la IA, incluidos los sectores público y privado, al sentar las bases

[824] Organización para la Cooperación y el Desarrollo Económicos, OCDE (2019). *Recommendation of the Council on Artificial Intelligence.* Disponible en: <https://legalinstruments.oecd.org/en/instruments/oecd-legal-0449>.

[825] G20. (2019). *G20 Ministerial Statement on Trade and Digital Economy*. Disponible en: <https://www.mofa.go.jp/files/000486596.pdf>.

para una evaluación del impacto ético de los sistemas de IA a lo largo de su ciclo de vida[826] (UNESCO, 2019).

### 20.2.4 Unión Europea (UE)

La UE también ha desempeñado un papel destacado en la regulación de la IA por lo que no podemos dejar de mencionar uno de los más recientes instrumentos normativos que se han presentado para regularla y menos considerando que este bloque económico tiene la intención de aumentar la inversión pública y privada durante la próxima década al menos 20 000 millones de euros al año para 2030. Sin duda, la labor que ha emprendido la Unión Europea al presentar la propuesta de la "Ley de Inteligencia Artificial"[827] (CNN Business, 2023) lidera el camino hacia la regulación tecnológica.

### 20.2.5 Reunión de los Ministros Digitales y Tecnológicos del G7

En abril se emitió la declaración ministerial sobre los retos actuales y futuros en la sociedad digital por el G7[828] (G7, 2023), donde se reafirmó el compromiso plasmado en la Declaración de la Cumbre por la Democracia, que aborda tecnologías emergentes como la IA, las biotecnologías, las tecnologías cuánticas y señala que deben ser moldeadas en línea con los principios democráticos, destacando la importancia de los debates internacionales sobre la interoperabilidad entre los marcos de gobernanza de la IA, reconociendo que las ideas afines pueden lograr una visión común entre los miembros del este grupo y así elaborar herramientas para una IA confiable, bajo marcos y estándares que puedan promover confiabilidad y permitir la evaluación de la IA.

### 20.2.6 España

Un ejercicio preventivo implementado por el Gobierno Español es un Sandbox[829] (2022), para garantizar el desarrollo de una Inteligencia

[826] Organización de Naciones Unidad para la Educación, la Ciencia y la Cultura (UNESCO), (2019). *Recomendación sobre la Ética de la Inteligencia Artificial.* Disponible en: <https://unesdoc.unesco.org/ark:/48223/pf0000380455_spa>.

[827] Ziady, Hanna. (2023). Europe is leading the race to regulate AI. Here's what you need to know. *CNN Business.* <https://edition.cnn.com/2023/06/15/tech/ai-act-europe-key-takeaways/index.html>.

[828] G7. (2023). *Ministerial Declaration The G7 Digital and Tech Minister's Meeting 30 de Abril 2023.* Disponible en: <http://www.g7.utoronto.ca/ict/2023-ministerial_declaration_dtmm.pdf>.

[829] Sandbox de regulación de IA en la UE. (2022). Disponible en: <https://portal.mineco.gob.es/RecursosNoticia/mineco/prensa/noticias/2022/20220627-Resumen-Piloto-Sandbox_IA-Final-ES.pdf>.

Artificial responsable y mitigar los riesgos potenciales de esta tecnología para la salud, la seguridad y los derechos fundamentales. El propósito está encaminado a poner en práctica las obligaciones futuras del Reglamento de IA y otros materiales de apoyo preparados para el futuro.

Por otro lado, el Consejo de Ministros aprobó un Real Decreto por el que se aprueba el estatuto de la Agencia Española de Supervisión de la Inteligencia Artificial (AESIA),[830] fruto del trabajo conjunto del Ministerio de Hacienda y Función Pública y el Ministerio de Asuntos Económicos y Transformación Digital.

### 20.2.7 Región latinoamericana

Esta región no se ha quedado atrás en el despliegue de instrumentos enfocados a emitir orientación y guía, para el uso adecuado de la IA, dirigidas a su gobernanza y regulación. Así mismo, se han identificado sandboxes regulatorios que, si bien ya llevan varios años desarrollándose a nivel mundial, aún no existe definición concreta sobre su alcance. Sin perjuicio de lo anterior, podríamos señalar que estos instrumentos son posibles ya que existe un margen de maniobra dentro de la regulación vigente que permita explorar de manera flexible sin sanciones por el incumplimiento de una norma.

Algunos ejemplos de los avances en la región latinoamericana son:

### 20.2.8 Brasil

En diciembre de 2022[831] (LexLatin,2023), se entregó al Senado el informe final de la comisión de juristas constituida para proponer un proyecto de regulación de la IA en Brasil, esta propuesta incluye medidas de gobernanza, con rendición de cuentas en caso de violación de la ley, un requisito de transparencia en el uso de la IA y la garantía del respeto de los derechos fundamentales con la directriz de que los algoritmos no acentúan formas de discriminación.

[830] Ibarra, E. (2023) *Aprobado el Estatuto de la Agencia Española de Supervisión de la Inteligencia Artificial*. Disponible en: <https://www.linkedin.com/pulse/aprobado-el-estatuto-de-la-agencia-espa%C3%B1ola-ernesto-ibarra-s?utm_source=share&utm_medium=member_ios&utm_campaign=share_via>.

[831] Ferreira, M. (abril 2023) *Brasil presenta proyecto para regular el uso de la inteligencia artificial.* Lexlatin. Disponible en: <https://lexlatin.com/opinion/brasil-regular-uso-inteligencia-artificial>.

### 20.2.9 Uruguay

Uruguay ha sido uno de los primeros países en ratificar su intención de implementar la Recomendación de UNESCO, de allí que será también de los primeros países de la región en aplicar la Metodología de Evaluación de Preparación (RAM, por sus siglas en inglés) de la UNESCO. Además, Actualmente, el gobierno uruguayo, a través de Agencia de Gobierno Electrónico y Sociedad de la Información y del Conocimiento (AGESIC) está actualizando su estrategia nacional de IA, con el fin de proponer acciones más relevantes y efectivas tras el avance en materia de IA[832] (UNESCO, 2023).

### 20.2.10 Perú

Pese a ser un país con grandes brechas digitales entre la población, Perú ha dedicado varios esfuerzos para fomentar el avance de la transformación digital en el sector público. En ese sentido, publicó la "Ley que promueve el uso de la Inteligencia Artificial (IA) en favor del desarrollo económico y social del país"[833] (Diario Oficial del Bicentenario "El Peruano", 2023), uno de sus primeros pasos específicos a favor de esta tecnología.

### 20.2.11 Chile

A través de su política de IA, Chile ha decidido avanzar en la experimentación regulatoria frente a los retos que estos sistemas pueden generar, permitiendo que, a su vez, este ejercicio se convierta en una fuente de conocimiento y experiencia, con la finalidad de motivar la innovación en el país y generar condiciones para un despliegue ambicioso de la tecnología en distintos sectores productivos.

La puesta en marcha de los sandboxes[834] (CAF, 2021) es una de las propuestas más ambiciosas de la Política de IA en Chile y el reto está en lograr su adecuada implementación, con la finalidad de promover

[832] UNESCO (junio 2023) *Inteligencia Artificial, Ciudadanía Digital y Educación.* Disponible en: <https://www.unesco.org/es/articles/inteligencia-artificial-ciudadania-digital-y-educacion>.

[833] Diario Oficial del Bicentenario El Peruano (julio, 2023), *Ley que promueve el uso de la Inteligencia Artificial en favor del Desarrollo Económico y Social del País.* Disponible en: <https://busquedas.elperuano.pe/normaslegales/ley-que-promueve-el-uso-de-la-inteligencia-artificial-en-fav-ley-n-31814-2192926-1/>.

[834] Banco de Desarrollo de América Latina (CAF), (agosto 2021) *Sandbox Regulatorio de Inteligencia Artificial en Chile* Disponible en: <https://www.economia.gob.cl/wp-content/uploads/2021/09/PaperSandboxIA.pdf>.

la discusión en la materia generando un mayor entendimiento sobre los sandboxes regulatorios y el impacto que están teniendo a nivel mundial. Chile también apuesta a ser competitivo en un escenario regional e internacional e irá adaptando las regulaciones existentes de esta tecnología.

### 20.2.12 Colombia

En Colombia se creó un instrumento a manera de recomendaciones con el objetivo de adoptar un marco ético como guía para la implementación de la inteligencia artificial en el sector público nacional, el objetivo de este marco[835] (Marco Ético para la Inteligencia Artificial en Colombia, 2021), es reconocer plenamente la necesidad de proteger y reforzar todos los derechos humanos de los ciudadanos en el desarrollo, el uso y la gobernanza de la IA, garantizando el respeto y la aplicación de los mismos.

### 20.2.13 República Dominicana

La región del Caribe también ha mostrado su interés y de acuerdo con el Government AI Readiness Index 2022, República Dominicana ocupa la novena posición como uno de los países más preaparados en materia de IA. Con el apoyo del Banco de Desarrollo de América Latina y El Caribe (CAF) y la UNESCO ha establecido una Agenda Digital 2030 y una Política Nacional de Innovación 2030, que ha establecido el mandato de desarrollar una Estrategia Nacional de IA en el 2023 (ENIA)[836] (CAF, 2023).

### 20.2.14 México

Por lo que hace a los esfuerzos de México, en 2023 se presentó la Alianza Nacional de Inteligencia Artificial (ANIA) que reunió voces de expertos, y autoridades interesadas con el objetivo de fortalecer el entorno de la IA en México a través de una conciencia colectiva basada en derechos humanos, principios de colaboración

[835] Marco Ético para la Inteligencia Artificial en Colombia, (Octubre, 2021), Disponible en: <https://dapre.presidencia.gov.co/TD/MARCO-ETICO-PARA-LA-INTELIGENCIA-ARTIFICIAL-EN-COLOMBIA-2021.pdf>.

[836] CAF, 2023. *Impulsando la inteligencia artificial en América Latina y el Caribe lecciones desde República Dominicana y Uruguay*. Disponible en: <https://www.caf.com/es/actualidad/noticias/2023/07/impulsando-la-inteligencia-artificial-en-america-latina-y-el-caribe-lecciones-desde-republica-dominicana-y-uruguay/>.

e interoperabilidad con una perspectiva ética, incluyente, integral, objetiva y multidisciplinaria.

Para garantizar una sociedad incluyente, participativa y que promueva la defensa y garantía de nuestros derechos humanos frente a las innovaciones tecnológicas, el INAI contribuirá haciendo de la protección de datos personales y la privacidad un marco vanguardista que sirva de referencia en el quehacer gubernamental y en el sector privado, mediante el cual se privilegie la efectiva tutela y garantía de estos derechos.[837] (Mendoza, J., 2023).

El Open Loop, capítulo de México[838] (Del Pozo, et al, 2020), es otro ejercicio coordinado y dirigido por Meta antes Facebook que desarrolló un Prototipo de Políticas Públicas (PPP) enfocado en el tema de transparencia y explicabilidad de los sistemas de IA para agregar más valor a sus usuarios y público en general que tendrán como resultado un informe de recomendaciones de política pública para los reguladores mexicanos.

Algunos ejemplos de los avances de Norteamérica son:

### 20.2.15 Estados Unidos de América (EUA)

En EUA se han presentado algunos avances como la propuesta de la “Ley de Responsabilidad Algorítmica”[839] (2022), que aborda los impactos de los sistemas automatizados y crea una nueva transparencia. También fue presentada la “Declaración de Derechos de la IA”[840] (2022), que es una guía para proteger a los ciudadanos frente a los sesgos y tratos desiguales que pueden generar los datos y el Instituto Nacional de Estándares y Tecnología (NIST) publicó el “Marco de Gestión de Riesgos de IA y lanzó su Centro de Recursos

[837] Mendoza, J. (abril, 2023) *El compromiso del Inai con la Alianza Nacional de Inteligencia Artificial.* El economista Disponible en: <https://www.eleconomista.com.mx/opinion/El-compromiso-del-Inai-con-la-Alianza-Nacional-de-Inteligencia-Artificial-20230423-0004.html>.

[838] Del Pozo, C. et al. (agosto 2023). *Open Loop Mexico: Public Policy Prototype on the Transparency and Explainability of Artificial Intelligence Systems*. Disponible en: <https://openloop.org/programs/ai-transparency-explainability-mexico/>.

[839] Ley de Responsabilidad Algorítmica, 2022. Disponible en: <https://www.wyden.senate.gov/imo/media/doc/Algorithmic%20Accountability%20Act%20of%202022%20Bill%20Text.pdf>.

[840] Declaración de Derechos de la IA, 2022. Disponible en: <https://www.whitehouse.gov/ostp/ai-bill-of-rights/>.

de IA”[841] (2023). Hasta ahora podría ser el marco más detallado de su tipo en EUA.

### 20.2.16 Canadá

En junio de 2022 el gobierno de Canadá lanzó la propuesta “Ley de Datos e Inteligencia Artificial (AIDA)[842]” (2022) como parte del Proyecto de Ley C-27, la Ley de Implementación de la Carta Digital de 2022. La AIDA representa un hito importante en la implementación de la Carta Digital y garantiza a los usuarios confiar en las tecnologías digitales que utilizan todos los días.

## 20.3 Gobernanza de la IA

Los gobiernos deberían adoptar un marco regulador que establezca un procedimiento para que las autoridades públicas, en particular, lleven a cabo evaluaciones del impacto ético de los sistemas de IA a fin de anticipar las repercusiones, atenuar los riesgos y establecer mecanismos de supervisión adecuados como la auditabilidad, la trazabilidad y la explicabilidad, que permitan evaluar los algoritmos, los datos y los procesos de concepción.[843] (UNESCO, 2019).

Las tecnologías disruptivas como la inteligencia artificial deben ser concebidas desde su diseño con valores determinados que respeten los derechos humanos, la democracia y la diversidad, ya que pueden amplificar las desigualdades y causar daños, en particular a los grupos vulnerables y marginados. Así mismo deben contemplar medidas de seguridad adecuadas en beneficio de las personas, mediante mecanismos de supervisión y cumplimiento, alineados a principios y reglas que permitan la rendición de cuentas en cualquier situación.

Nuestras sociedades deben adaptarse a la transformación que traerá la IA a través de cambios en su marco de cooperación y modelo de gobernanza. La construcción de una sociedad inteligente centrada

[841] Marco de Gestión de Riesgos de IA, 2023. Disponible en: <https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf>.

[842] Ley de Datos e Inteligencia Artificial (AIDA), 2022. Disponible en: <https://ised-isde.canada.ca/site/innovation-better-canada/en/artificial-intelligence-and-data-act-aida-companion-document>.

[843] Organización de Naciones Unidad para la Educación, la Ciencia y la Cultura (UNESCO), (2019). Recomendación sobre la Ética de la Inteligencia Artificial. Disponible en: <https://unesdoc.unesco.org/ark:/48223/pf0000380455_spa>.

en el ser humano requiere la plena cooperación del gobierno, las empresas, las organizaciones sociales y la academia. El control humano continuo es fundamental para garantizar que los algoritmos no conduzcan a resultados no deseados o no controlados.

### 20.3.1 Principios de Transparencia y rendición de cuentas

Una IA confiable depende de la rendición de cuentas, lo que a su vez presupone que debe existir transparencia y explicabilidad en los sistemas relacionados con esta tecnología durante todo su ciclo de vida. La transparencia refleja el grado en que la información sobre un sistema de IA y sus resultados están disponibles para las personas que interactúan con dicho sistema, independientemente de si son siquiera conscientes de que lo están haciendo.

La transparencia y la explicabilidad hacen que las personas conozcan y comprendan la forma en cómo se recolecta y procesa la información y los fines de hacerlo, en especial, cuando se tratan datos personales. Sin embargo, al día de hoy, existe una llamada caja negra en la toma de decisiones de los sistemas de IA, esto quiere decir que existen procesos en el funcionamiento del sistema que no se sabe de dónde surgen. Por lo que es indispensable sumar esfuerzos para identificar uno a uno los motivos por los que el sistema llegó a una decisión, ya que el resultado puede afectar los derechos, las libertades e incluso la seguridad del individuo. Por lo tanto, Los sistemas de IA deben ser explicables, la información debe ser de fácil acceso y comprensible para así poder evaluar su impacto. A una mayor transparencia, una mejor comprensión del funcionamiento de estos sistemas.

Para hacer posible la rendición de cuentas, debería considerarse la implementación de mecanismos adecuados de supervisión, evaluación del impacto, auditoría y diligencia debida, incluso en lo que se refiere a la protección de los denunciantes de irregularidades, para garantizar la rendición de cuentas a lo largo de su ciclo de vida.

Los desarrolladores y las autoridades de control son responsables de garantizar la auditabilidad y la trazabilidad del funcionamiento de los sistemas de IA, lo anterior para que de manera preventiva se pueda dar atención y solución a los posibles conflictos que pudieran surgir con relación a las posibles vulneración y amenazas de los derechos humanos.

## 20.4 Panorama Latinoamericano de la IA

Recientemente, fue presentado el primer índice Latinoamericano de Inteligencia Artificial (ILIA), que contempla el análisis exhaustivo de la situación de la IA en 12 países de América Latina (Argentina, Bolivia, Brasil, Chile, Colombia, Costa Rica, Ecuador, México, Panamá, Paraguay, Perú y Uruguay). Si bien este estudio no contempla la totalidad de países que integran esta región, destaca los esfuerzos que se han realizado en torno a la implementación de sistemas que conllevan el uso de la IA.

ILIA es un estudio público que ofrece una visión detallada y amplia del estado actual de la IA en la región de América Latina y el Caribe. Con una mirada centrada en la pertinencia local, el estudio cubre en detalle temáticas como infraestructura, capital humano, disponibilidad de datos, regulaciones, áreas estratégicas y participación ciudadana, entre otros.

Uno de los principales hallazgos que revela este estudio es que los países que cuentan con una Estrategia Nacional de IA vigente muestran mayor inserción institucional y un desarrollo regulatorio más armónico, tanto en esta dimensión como en el ILIA en general. Si bien no es una relación causal, contar con una estrategia consensuada a nivel local parece ser un punto de partida para otros elementos respecto al ambiente institucional.

Asimismo, Argentina, Brasil y México, son referentes regionales en cuanto a la participación en espacios internacionales para incidir en la discusión global sobre IA. Chile muestra el mejor desempeño en cuanto a participación ciudadana en la formulación de estrategias. Perú destaca en materia de regulación y legislación relacionada a la IA o a la protección de datos. Sin embargo, cada país puede mejorar en al menos, una subdimensión de gobernanza.

Los mejores puntajes del índice los obtuvieron Chile (72.67) y Brasil (65.31) quienes se erigen como los líderes en la región. Por su parte, México (48.55) cuenta con un ecosistema maduro y sólido, con buen desempeño en investigación, pero aún hace falta fortalecer la infraestructura y la formación profesional, e impulsar la innovación y el desarrollo.

## 20.5 Propuesta latinoamericana hacia la creación de un mecanismo exprofeso que coadyuve en asuntos relacionados con la IA.

Derivado del análisis de los resultados del índice Latinoamericano de Inteligencia Artificial, se han detectado áreas y oportunidades de mejora en nuestra región. Por un lado, debemos garantizar que la brecha digital existente disminuya lo más posible entre naciones y entre las poblaciones y garantizar el acceso equitativo a internet y a los sistemas de inteligencia artificial. Por otro lado, se deberá considerar que será responsabilidad de las autoridades competentes en cada jurisdicción sentar las bases y recomendaciones para el uso adecuado de esta tecnología, poniendo en marcha acciones de difusión, concientización y capacitación para la población en general, lo que nos permitirá llevar a buen término la adopción de una regulación en la materia sin afectar o detener la innovación tecnológica, que cada día ofrece más opciones que facilitan la vida de los usuarios.

Para que lo anterior suceda, y con la finalidad de homologar o establecer los estándares mínimos para la regulación de la IA en la región, se propone crear un Comité de Expertos integrado por especialistas multidisciplinarios que colaborarán con la finalidad de investigar, dar atención y seguimiento a los casos que surjan a partir de las vulneraciones causadas por estas tecnologías en contra de los derechos humanos de protección de datos personales y privacidad de los usuarios.

La ruta a seguir para impulsar esta propuesta sería bajo el esquema que ha recorrido la presentación y promoción de la resolución CJI/doc. 673/22 rev.1, relativa al proyecto de principios interamericanos en materia de neurociencias, neurotecnologías y derechos humanos, presentado por el Dr. Ramiro Orias Arredondo, Miembro del Comité Jurídico Interamericano y relator del tema ante la Organización de Estados Americanos (OEA).

Otra vía, sería contar con el apoyo de la persona designada para la relatoría especial para la protección de datos personales de la OEA, como ocurrió con la aprobación de los Principios Actualizados sobre la Privacidad y la Protección de Datos Personales elaborados por el Comité Jurídico Interamericano, mediante resolución AG/RES. 2975 (LI-O/21), en noviembre de 2021.

Este marco normativo referencial nos permitirá encontrar el camino adecuado para que la propuesta tenga el alcance esperado y sea del interés de los miembros del Comité Jurídico Interamericano para posicionarlo ante los miembros que integran la Asamblea General de esta Organización.

Como se señalaba, para facilitar la implementación de esta propuesta, es deseable la creación de un Comité de Expertos que previamente analice y consensúe la importancia y urgente necesidad de coadyuvar, a través de mecanismos no vinculantes, la situación en torno al uso e implementación de las tecnologías disruptivas existentes y por desarrollar, dado el riesgo que podrían implicar en la vida privada de los usuarios.

Este Comité de Expertos, además, deberá contar con atribuciones para dirigir evaluaciones de impacto en materia de derechos humanos y evaluaciones de impacto éticas respecto al uso y alcance de las tecnologías disruptivas. El comité deberá estar integrado por múltiples partes interesadas provenientes de diferentes sectores y con perfiles multidisciplinarios que permitan el análisis objetivo y técnico de cada uno de los casos específicos que se aborden no solo bajo una óptica de privacidad sino integral.

Una vez conformado y con una propuesta clara sobre el rumbo de esta iniciativa debemos allanar el camino para contar con el apoyo de las instancias dentro de la OEA para presentarla ante la Asamblea General y, de considerarse viable, por medio de una resolución se mandate al Comité Jurídico Interamericano para integrar dicho Comité Expertos y se establezcan las reglas de operación de la designación de sus integrantes y de los casos a los que se daría seguimiento y atención.

El Comité en principio estaría integrado por expertos en privacidad, protección de datos personales y desarrolladores tecnológicos tanto del sector privado como público, investigadores, representantes de la academia, filósofos, sociólogos, representantes de redes regionales e internacionales y organizaciones de la sociedad civil. El procedimiento de selección sería por dos vías, a través de una convocatoria pública y mediante invitación directa a perfiles que por su experiencia podrían formar parte de este grupo selecto. Una vez elegidos, se trabajará el contenido de una resolución y las posibles acciones a implementar para ampliar el alcance y difusión de los casos en cuestión y con ello formalizar la propuesta a través del Comité Jurídico al resto de miembros que integran la Asamblea General.

El objetivo de este Comité de Expertos deberá estar basado en la buena voluntad y se fundamentará en el intercambio de conocimientos y buenas prácticas que fomenten la cooperación internacional, con base en el multilateralismo y las oportunidades que este nos ofrece para fortalecer la tutela de los derechos humanos, sumando esfuerzos con otros organismos internacionales que también se han pronunciado al respecto, así como con los grupos de las potencias económicas que han mostrado su preocupación ante este panorama de la nueva era digital.

Para garantizar que esta propuesta sea incluyente y tome en cuenta diversas perspectivas se buscará realizar procesos consultivos que resulten en el apoyo y respaldo político de las partes involucradas para obtener el alcance deseado. El trabajo de este Comité estaría basado en un mecanismo que buscará:

- Analizar casos específicos

  Los expertos y expertas de este mecanismo, con el apoyo de la Asamblea General, analizarán las leyes e instituciones nacionales y evaluarán si éstas son efectivas para la prevención de vulneraciones a los derechos humanos en función de la tecnología de que se trate.

- Emitir recomendaciones

  El Comité por mandato de la resolución, formulará recomendaciones a los Estados para que mejoren y adecuen sus marcos jurídicos e instituciones, tomando en consideración los principios incluidos en la normatividad existente de derecho blando (UNESCO, OCDE) en términos de inteligencia artificial como la rendición de cuentas, la transparencia y la explicabilidad.

- Brindar seguimiento

  El Comité dará seguimiento a las recomendaciones hechas a cada Estado para evaluar su implementación y los avances logrados por cada país en la materia.

- Elaborar herramientas de cooperación

  Siendo la cooperación uno de sus pilares, el Comité elaborará herramientas como leyes modelo, principios y guías legislativas para que los Estados cuenten con ellas a la hora de llevar adelante reformas en sus marcos jurídicos con el fin de fortalecerlos para prevenir vulneraciones a los derechos humanos como la privacidad y la protección de datos personales derivados del uso de tecnologías.

# 21 Aspectos normativos de la IA en Argentina

**María Julia Giorgelli, Experta independiente.**

## Resumen

El documento reseña el entramado normativo sobre la inteligencia artificial (en adelante IA) en Argentina. Asimismo, brinda información de contexto sobre los compromisos internacionales refrendados por el país; enumera diversas acciones llevadas adelante por el Poder Ejecutivo Nacional y resume los últimos proyectos de ley presentados a nivel nacional. En todos los casos se pone el eje en el derecho a la privacidad/datos personales y la transparencia/información.

## 21.1 Introducción

En el año 2022, el Instituto de Estadísticas y Censo de Argentina relevó que el 62,6% de los hogares urbanos tienen acceso a computadora y el 92,1% a Internet. Además, informó que 89 de cada 100 personas usan teléfono celular y 88 de cada 100 utilizan Internet.[844]

Diariamente en ámbitos como el trabajo, la educación, la seguridad, la administración de justicia, la salud o el intercambio de bienes y servicios se usan diversos sistemas basados en inteligencia artificial. También se lo hace en sectores menos vitales, aunque igualmente relevantes como es el caso del entretenimiento. A la par que el fenómeno crece vertiginosamente centrado en la eficiencia y productividad que agregan, se visibilizan dificultades que impactan en los derechos fundamentales de las personas.

Diversos sectores y actores alzaron su voz evidenciando los problemas que acarrea. Entre ellas, las numerosas críticas del filósofo francés Eric Sadin[845] o también las apreciaciones de Kate Crawford[846] quien afirma que la IA no es "ni artificial ni inteligente". En la misma línea, la obra colectiva argentina "Pensar la tecnología digital con

[844] Encuesta Permanente de Hogares (EPH) Acceso y uso de tecnologías de la información y la comunicación. Cuarto trimestre de 2022 <https://www.indec.gob.ar/uploads/informesdeprensa/mautic_05_239BB78E7691.pdf>.

[845] Sadin, E. (2020). *LA INTELIGENCIA ARTIFICIAL O EL DESAFIO DEL SIGLO.* Caja Negra.

[846] Crawford, K. (2022). *Atlas de la Inteligencia Artificial: poder, politíca y costos planetarios.* Fondo de Cultura Economica.

perspectiva de género" sostiene que con estos avances se consolidó el statu quo, con situaciones de inequidad e injusticia por parte de los actores poderosos en detrimento de sectores marginalizados o minimizados.[847]

Estamos ante un producto complejo y global que responde a su tiempo y es evidente que requiere de un abordaje multidisciplinario. Además, debería poder reflejar las particularidades de cada región, dotar de garantías a los afectados y establecer ciertas salvaguardas como las pruebas previas, la supervisión humana, la obligación de informar y transparentar el procesamiento y utilización de los datos y, llegado el caso, responsabilizarse ante eventuales daños.

En síntesis, corresponderá trabajar para que efectivamente los sistemas de IA estén centrados en el bien público y no en el mercado y la productividad.[848]

## 21.2 Situación normativa en Argentina.

### 21.2.1 Normas generales

Al día de hoy, Argentina no cuenta con una regulación específica en materia de inteligencia artificial. Sin embargo, hay un contexto normativo que se relaciona y favorece la temática. Concretamente, existen directrices supranacionales y leyes (previas al fenómeno de la IA) que garantizan derechos off-line y de promoción del sector que contribuyen a enmarcar la cuestión.

A nivel supranacional encontramos dos precedentes gravitantes que fueron reconocidos por nuestro país. Ambos constituyen "softlaw" y están dirigidos centralmente a los Estados nación; destacan la complejidad del fenómeno que enfrenta la humanidad y acuerdan en la necesidad del trabajo interdisciplinario, así como que la tecnología tiene que estar centrada en el bien común. Son principios que se repetirán en diversos compendios y constituyen la columna vertebral de la cuestión.

[847] Balmaceda, T., Pedace, K., & Lawler, D. (2021). *Pensar la tecnologia con perspectiva de género* <https://proyectoguia.lat/wp-content/uploads/2022/06/perspectiva-generoV6.pdf>.

[848] Guerra, J. (2023). *Hacia un marco feminista para el desarrollo de IA: de los princpios a la práctica.* America Latina: Digital Creative Commons Atribución 4.0 Internacional.

En los "Principios sobre la Inteligencia Artificial" suscripto por treinta y ocho países y ocho observadores[849] se establece un compromiso que aspira a lograr sistemas de IA robustos, seguros, imparciales y fiables. En dichos Principios se incluye una previsión específica sobre *"transparencia y explicabilidad"*[850] que apunta a quien hubiere sido afectado a comprender de manera sencilla la lógica que sirvió de base para la predicción que lo dañó. La protección de los datos personales es un concepto que aparece en distintos pasajes de la norma, pero sin autonomía. A pesar de ello, sí se subraya la necesidad de un desarrollo de la IA libre de sesgos.

El segundo precedente lo constituye la "Recomendación sobre la ética de la Inteligencia Artificial" emitidas por la UNESCO en el año 2021. Es un compendio clave por ser el primer documento global. En esa oportunidad se consideró el valor ético como el nodo de la norma subrayando la importancia de respetar la dignidad humana y proteger el bien común. La directriz es más densa y amplia que la anterior, seguramente porque recoge el debate que se dio en esos tiempos. Hace hincapié en aspectos relacionados con el medioambiente, diversidad, inclusión y no discriminación. En dichas Recomendaciones, existe un apartado particularizado sobre "protección del derecho a la privacidad y los datos personales" así como también sobre "información y transparencia".[851]

Respecto a los datos personales y la privacidad la norma se alinea con una concepción continental-europea del derecho, por lo que posee sumo valor para nuestro país que sostiene un esquema análogo. Apela a que se garantice la protección de la información personal durante todo el ciclo de vida del sistema de IA. También se menciona la necesidad de realizar estudios de impacto a la privacidad, herramienta que permitirá, mediante una evaluación previa, evaluar el uso de

[849] Conforme <https://www.oecd.org/acerca/miembros-y-socios/>.

[850] Apartado 1.3 *"Los actores de la IA deben comprometerse con la transparencia y la divulgación responsable de los sistemas de IA. Para ello deberán proporcionar información significativa, adecuada al contexto y coherente con el estado de la técnica: i. Fomentar una comprensión general de los sistemas de IA. ii. concienciar a las partes interesadas sobre sus interacciones con los sistemas de IA, incluso en el lugar de trabajo, III. para permitir que los afectados por un sistema de IA comprendan el resultado y, IV. para permitir que aquellos afectados negativamente por un sistema de IA cuestionen su resultado basándose en información sencilla y fácil de entender sobre los factores y la lógica que sirvió de base para la predicción, recomendación o decisión"*; disponible en <https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449>.

[851] Derecho a la intimidad y protección de datos apartados 32, 33 y 34. Transparencia y Explicabilidad apartado 37 y 38, disponible en <https://unesdoc.unesco.org/ark:/48223/pf0000381137_spa>.

dichos sistemas. En relación con la información que debe brindarse a las personas, afirma la necesidad de *"aumentar la transparencia y la aplicabilidad de los sistemas de IA"* y agrega la idea de que ello sea "adecuado al contexto".

Como se mencionó, en el sistema jurídico interno existe legislación que garantiza ciertos derechos y otra, relacionada, que identifica a la "tecnología" como un valor a jerarquizar y proteger.

La propia Constitución Nacional señala como atribución del Congreso Nacional la de "*Proveer lo conducente al desarrollo humano, al progreso económico con justicia social, a la productividad de la economía nacional, a la generación de empleo, a la formación profesional de los trabajadores, a la defensa del valor de la moneda, a la investigación y al desarrollo científico y tecnológico, su difusión y aprovechamiento".*[852] En análoga dirección se identifican la siguiente normativa: "Ciencia, tecnología e innovación" (Ley Nro. 25467 del año 2001[853]), el "Régimen de Promoción de la Economía del Conocimiento" (Ley Nro. 27506 del año 2019[854]) o la más reciente sobre "Financiamiento del sistema nacional de ciencia, tecnología en innovación" (Ley Nro. 27614 del año 2021[855]).

Mención aparte requiere la legislación sobre "Protección de datos personales". Por un lado, porque se trata de un derecho que tiene jerarquía constitucional y también porque la Ley Nro. 25326 del año 2000 prevé una garantía específica que entrelaza la protección de los datos personales y la transparencia como derechos expresos, posibles de reclamar ante quien trate nuestra información personal (Capítulo III, art. 13 y cc[856]). También existe un artículo específico sobre tratamiento automatizado de datos, aunque respecto a decisiones judiciales o los actos administrativos (art. 20). En este punto vale recalcar que dicha ley se encuentra en proceso de actualización

[852] Constitución de la Nación Argentina <https://servicios.infoleg.gob.ar/infolegInternet/anexos/0-4999/804/norma.htm>.

[853] Disponible en http://servicios.infoleg.gob.ar/infolegInternet/anexos/65000-69999/69045/texact.htm>.

[854] Disponible en http://servicios.infoleg.gob.ar/infolegInternet/anexos/320000-324999/324101/texact.htm>.

[855] Disponible en http://servicios.infoleg.gob.ar/infolegInternet/anexos/345000-349999/347804/norma.htm>.

[856] Disponible en <https://servicios.infoleg.gob.ar/infolegInternet/anexos/60000-64999/64790/texact.htm>.

estableciendo el proyecto su aplicación con independencia de las técnicas, procesos o tecnologías –actuales o futuras- que surjan.

Asimismo, serían de aplicación las previsiones del “Convenio n° 108 del Consejo de Europa para la protección de las personas con respecto al tratamiento automatizado de datos de carácter personal y protocolo adicional al convenio para la protección de las personas con respecto al tratamiento automatizado de datos de carácter personal, a las autoridades de control y a los flujos trasfronterizos de datos conocido como Convenio 108 (Ley Nro. 27.483 del año 2019) junto a su versión modernizada.

En cuanto a la estructura orgánica estatal se destaca la existencia del Ministerio de Ciencia, Tecnología e Innovación[857] dentro del ámbito del Poder Ejecutivo Nacional. También se señala otro foco con competencia en la temática, pero esta vez dependiente de la Jefatura de Gabinete de Ministros. Desde este último organismo, en junio del corriente, se emitieron las “Recomendaciones para una Inteligencia Artificial Fiable”.[858] Se trata de una norma orientadora, no obligatoria, pero muy interesante dado que arroja pautas para los proyectos de innovación del sector público. La misma desarrolla un marco conceptual, describe los diversos ciclos de la IA, y prevé recomendaciones y directrices para implementar un proyecto de IA por lo que puede servir como una guía práctica frente a proyectos de innovación.

Además, es la primera directriz estatal formal. Entre sus aspectos nodales garantiza la protección del derecho a la intimidad y protección de datos junto a la transparencia y explicabilidad desde la etapa del diseño de las herramientas, en línea con las previsiones supranacionales. Desde la Jefatura de Gabinete de Ministros por Decisión Administrativa 750/2023 y con claros fines de coordinación de la temática, se crea la MesaInterministerial sobre Inteligencia

[857] En el curso de este año el Gobierno nacional anunció un incremento en el presupuesto 2023 para el área ciencia y técnica. Concretamente se trata de un aumento en relación al año anterior de un 22,69% <https://chequeado.com/el-explicador/ciencia-como-evoluciono-la-inversion-en-la-gestion-de-alberto-fernandez/>. Asimismo se anunció la presentación ante el Congreso Nacional del “Plan Nacional de Ciencia y Tecnología 2030” prevé la tecnología de la IA como un aspecto en el cual trabajar <https://www.argentina.gob.ar/sites/default/files/plan_nacional_de_cti_2030.pdf>.

[858] Jefatura de Gabinete de Ministros, A. (2023). Disposición 2 / 2023 Recomendaciones para una Inteligencia Artificial Fiable <https://www.argentina.gob.ar/sites/default/files/2023/06/recomendaciones_para_una_inteligencia_artificial_fiable.pdf>.

Artificial con el fin de abordar el avance y aplicación de la Inteligencia Artificial en diversos sectores de la economía y de la sociedad, de conformidad con un marco ético, de desarrollo sostenible y de transformación digital, y con la finalidad de diseñar una estrategia integral al respecto para ser aplicada por el Poder Ejecutivo Nacional.

Asimismo, el país cuenta con un órgano administrativo denominado Agencia de Acceso a la Información Pública[859] que tiene por objeto garantizar el cumplimiento de los derechos de acceso a la información pública y la protección de datos personales. Desde dicha área recientemente se creó un "Programa de transparencia y protección de datos personales en el uso de la Inteligencia Artificial",[860] entre sus objetivos se prevé: indagar acerca de las implicancias sociales, económicas, laborales, culturales y ambientales del desarrollo de la IA en Argentina; analizar el estado de situación respecto al uso de la IA en las organizaciones del Sector Público nacional; generar conocimiento que permita identificar buenas prácticas, aprendizajes y recomendaciones en materia de transparencia, transparencia algorítmica y protección de datos personales en el uso de la IA; realizar acciones de fortalecimiento de capacidades institucionales de actores claves en la implementación de la IA en materia de transparencia y protección de datos personales; impulsar procesos participativos para la generación de propuestas regulatorias de la IA en el país, en coordinación con otros organismos gubernamentales competentes en la materia. Atento su reciente creación aún no tiene ejecución que permita habilitar comentarios al respecto.

Otro dato relevante son diversas medidas promovidas por el Estado que dan cuenta de que el tema está en agenda. Así, a nivel de la administración pública central se han realizado una serie de foros y emitido declaraciones que sin duda establecen un andamiaje y suman al debate.[861] En la misma dirección se han realizado nuevos anuncios, como el financiamiento por un valor de USD 35.000.000 de dólares destinado al Programa de Apoyo a las Exportaciones

[859] Sitio oficial de la AAIP <https://www.argentina.gob.ar/aaip>.

[860] Resolución 161/2023 disponible <https://www.boletinoficial.gob.ar/detalleAviso/primera/293363/20230904>.

[861] Por ejemplo <https://www.argentina.gob.ar/noticias/inteligencia-artificial-el-debate-principal-es-como-se-regula-y-democratiza y <https://www.argentina.gob.ar/noticias/se-realizo-el-encuentro-oportunidades-y-desafios-en-inteligencia-artificial-0>.

de la Economía del Conocimiento para el desarrollo del sector y su inserción internacional.[862]

Por último, para completar este escenario resulta interesante referirse al reciente "Índice Latinoamericano de Inteligencia Artificial".[863] Se trata de un estudio efectuado sobre doce países de la región con el apoyo de la CEPAL y presentado en agosto de este año en Chile. El mismo compara tres dimensiones: 1. Factores Habilitantes, 2. Investigación y desarrollo, 3. Gobernanza. Las conclusiones son positivas respecto de Argentina donde se destaca una sólida base para desarrollar y aprovechar el potencial de la IA. También se menciona la necesidad de generar regulaciones específicas en IA, al tiempo que se subraya positivamente la infraestructura en términos de conectividad y un alto potencial en desarrollo de talento y habilidades tecnológicas. En materia de "gobernanza" las calificaciones son altas.

### 21.2.2 Proyectos de Ley

A continuación, se reseñan tres propuestas normativas. El corte fue considerando las presentaciones hechas en el año por ante el Congreso Nacional por lo que conservan estado parlamentario.[864]

Un primer proyecto, Nro. 1472-D-2023[865] plantea la reforma de una ley vigente (Ley Nro. 25467 "Ciencia, tecnología e innovación") dicha propuesta es escueta, posee tres artículos y hace foco centralmente en aspectos generales basados en la ética. Igualmente es positivo porque prevé que todo avance en materia de IA deberá asegurar diversidad e inclusión, incluyendo la participación de todas las personas y grupos. Asimismo, señala que se deberá asegurar la paz y justicia. La propuesta posee buenas intenciones, aunque no está a la altura de la complejidad del fenómeno ya que no se observan herramientas concretas que permitan garantizar esas declaraciones

[862] Información disponible en <https://www.argentina.gob.ar/noticias/nuevo-programa-de-35-millones-de-dolares-para-el-desarrollo-de-la-inteligencia-artificial>.

[863] CEPAL. (2023). *Indice Latinoamericano de Inteligencia Artificial* <https://indicelatam.cl/wp-content/uploads/2023/08/CAP-G-ARGENTINA.pdf>.

[864] Son los expedientes 2505-D-2023, 2504-D-2023 y 1472-D-2023. También vale mencionar una presentación previa en el mismo Congreso Nacional del año 2019 con el fin de crear un Consejo Federal de Inteligencia Artificial (Expediente 0509-D-2019). También existe alguna iniciativa local como por ejemplo el proyecto de la Ciudad de Buenos Aires (Expediente 2093-2023) tendiente a establecer la capacitación obligatoria en datos e IA.

[865] Disponible en <https://www.hcdn.gob.ar/proyectos/resultados-buscador.html>.

y/o enfrentar otros innumerables problemas que puede plantear la IA como, por ejemplo, los eventuales daños.

Entre algunos de los casos más resonantes en el país citamos los derivados de detenciones arbitrarias mediante el uso del sistema de sistema de reconocimiento facial,[866] o las erradas predicciones de embarazos adolescentes en la provincia de Salta que según informó la Universidad de Buenos Aires se detectaron "serios errores técnicos y conceptuales, que ponen en duda los resultados reportados y comprometen el empleo de dicha herramienta, sobre todo tratándose de una cuestión tan sensible".[867] Finalmente, en el texto no hay menciones a cuestiones relacionadas con la protección de los datos personales, ni transparencia.

Los otros dos proyectos (2504-D-2023 y 2505-D-2023[868]) son propuestas normativas autónomas. Uno de ellos tiene por fin crear un marco legal para la regulación del desarrollo y uso de la IA y el otro regularla en el ámbito educativo. Ambos gozan de una estructura adecuada ya que prevén fundamentos, objeto, principios rectores y garantías. Incluyen también herramientas útiles como los "glosarios" o la necesidad de la "capacitación" que coadyuvan a concretar temáticas complejas como es el caso de la IA. Además, incluyen previsiones presupuestarias, cuestión de importancia para concretar las iniciativas.

El denominado "Marco legal para la regulación del desarrollo y uso de la inteligencia artificial" (Nro. 2505-D-2023) tiene una pretensión totalizadora en materia de IA. Está compuesto por treinta y tres artículos, dispone la creación de un órgano especializado en IA y dedica asimismo dos artículos donde expresamente hace referencia a la protección de los datos personales y la privacidad, y a garantizar el derecho de los usuarios a entender el funcionamiento de tales sistemas. En cuanto a privacidad y datos personales: *"Los sistemas de IA deben respetar y proteger la privacidad de los usuarios y el tratamiento de sus datos personales de acuerdo con la normativa*

[866] A modo de ejemplo se cita: <https://www.pagina12.com.ar/209910-seis-dias-arrestado-por-un-error-del-sistema-de-reconocimien>.

[867] Material de consulta disponible en <https://liaa.dc.uba.ar/es/sobre-la-prediccion-automatica-de-embarazos-adolescentes/>.

[868] Disponible en <https://www.hcdn.gob.ar/proyectos/resultados-buscador.html y <https://www.hcdn.gob.ar/proyectos/resultados-buscador.html>.

*vigente de protección de datos aplicables...”* (art. 6). Por su parte sobre transparencia y explicabilidad señala que: “*Se requiere que los sistemas de IA sean transparentes en su funcionamiento, de manera que los usuarios comprendan cómo se toman las decisiones y se llega a los resultados. Se establece el derecho de los individuos a solicitar explicaciones de las decisiones tomadas por sistemas de IA que les afecten”* (art. 7).

La otra propuesta legislativa denominada “Ley de Regulación y Uso de la Inteligencia Artificial en la Educación” (Nro. 2504-D-2023) dispone un artículo específico sobre privacidad y protección de datos personales, pero no lo hace respecto de transparencia de manera autónoma. Sobre privacidad y protección de datos personales señala: *“Protección de datos y privacidad a) Recopilación de datos: Las instituciones educativas y los proveedores de IA deben obtener el consentimiento informado de los estudiantes o sus tutores legales antes de recopilar y utilizar datos personales con fines educativos....c) Derechos de los estudiantes: Los estudiantes tienen derecho a acceder, corregir y eliminar sus datos personales, así como a solicitar la interrupción del uso de la IA en su educación.”* (art. 4).

> Evidentemente, los marcos normativos adecuados mejoran los derechos de las personas. Cabe preguntarnos, sin embargo, si sólo con ello evitarán los sesgos o nos permitirá oponernos a las predicciones negativas o erróneas. Crawford reflexiona sobre las prácticas del sector y, al analizar los sistemas de reconocimiento facial, sostiene que a las imágenes “*No se las ve como individuos, sino como parte de un recurso técnico compartido, son tan solo otro componente de datos del programa de pruebas de verificación de reconocimiento facial, la referencia por excelencia en este campo”* luego agrega que *“Las bases de datos presagian la irrupción de una lógica que ahora ha invadido al sector tecnológico: la creencia de cualquier cosa puede ser un dato y de que los datos están ahí para que los tome quien quiera. No importa dónde fue tomada la fotografía, o si refleja un momento de vulnerabilidad o de dolor, o si representa una forma de humillación para el sujeto. Tomar y usar*

> *lo que esté disponible se ha normalizado tanto en toda la industria que pocos se detienen a cuestionar las políticas subyacentes a esas acciones."*[869]

En resumen, sin lugar a dudas las previsiones legales expresas coadyuvarán la posición de los individuos para reclamar por sus derechos, sin embargo, no puede ser lo único. También ello debe ir acompañado de una serie de acciones como dotar de recursos a las autoridades competentes, realizar acciones de difusión o concientización, sostener políticas públicas pro activas en favor de las personas y llevar adelante acciones de cumplimiento o "enforcement" que permitan concretar cambios y correcciones de rumbos.

## 21.3 Conclusiones

Hay mucho por decir sobre el fenómeno de la IA.

Sabemos que es una herramienta que impacta en diversos campos como las relaciones sociales, la subjetividad, el derecho, el medioambiente o la economía. Aun destacando que su avance es positivo, es obvio que aún resta enfrentar diversos desafíos, como impedir la reproducción de sesgos, y los diversos daños y afectaciones a los derechos fundamentales.

Esas acciones no pueden quedar sólo en manos de la industria. Creer que dicho sector trabajará priorizando el bien común y la protección de las personas constituye una ilusión. Como sostiene el filósofo francés Eric Sadin, estos sistemas poseen una lógica colonialista,[870] razón por la cual es importante trabajar para representar las particularidades de los países que integramos el sur global.

Por ello es necesario trabajar con premura de manera interdisciplinaria para que sea realidad una IA centrada en el bien común.

[869] Crawford, K. (2022). *Atlas de la Inteligencia Artificial: poder, politíca y costos planetarios.* Fondo de Cultura Economica.

[870] Sadin, E. (2020). *LA INTELIGENCIA ARTIFICIAL O EL DESAFIO DEL SIGLO.* Caja Negra.

# 22 IA y neurotecnologías: necesidad de protección ante nuevas encrucijadas

**Natalia L. Monti, Fundación Kamanau, Abogada, Magíster en DDHH, integra el Centro de Protección de Datos Personales de la Defensoría del Pueblo CABA, Argentina.**

*A la fecha no existe un texto normativo que realice una revisión sobre la incidencia aplicada que puede tener la ciencia sobre la integridad física y psíquica del ser humano y cómo ella podría afectar su derecho a la vida y a su integridad física y psíquica, existiendo la necesidad de protección ante estas nuevas encrucijadas.*[871]

## Resumen

A propósito de la Declaración de Principios Interamericanos en materia de neurociencias, neurotecnologías y derechos humanos de la OEA, aprobada en marzo de 2023 y además del reciente fallo de la Corte Suprema chilena contra la empresa que fabrica y comercializa dispositivos de neurotecnologia, Guido Girardi c/ Emotiv, del agosto de 2023. Se observa la necesidad de establecer criterios regulatorios que pongan el claro los estándares internacionales sobre los nuevos avances de la ciencia y la tecnología, siempre con una perspectiva en base a los derechos humanos. Los desafíos serán generar escenarios confiables en el desarrollo de estas nuevas tecnologías, sobre todo las tecnologías inmersivas que utilizan neurotecnología e inteligencia artificial ante los sectores más vulnerables, como son los niños, niñas y adolescentes, géneros y disidencias y personas con discapacidad, entre otras.

## 22.1 Avances en el ámbito científico

En los últimos años el acelerado desarrollo de las nuevas tecnologías de la inteligencia artificial (IA) está generando muchas preocupaciones, sobre todo en relación a los criterios éticos y jurídicos en los que deben asentarse estas innovaciones de la ciencia.

[871] C. S. Girardi c/ Emotiv Inc., 105065-2023 (Tercera Sala agosto 9, 2023).

Tal es así que en relación a la protección de los derechos humanos se están promoviendo distintas iniciativas, debido al creciente impacto que tienen estos nuevos avances científicos no solamente sobre la vida de las sociedades, sino también por la afectación que pueden significar sobre la libertad, el pensamiento y la integridad física; generándose una nueva frontera, antes poco conocida para el mundo jurídico.

En este sentido, es sabido que determinadas aplicaciones permiten una conexión bidireccional entre el sistema nervioso central de un individuo y un sistema electrónico. Tal es así que aparece la posibilidad de acceder a los datos que arroja la información cerebral y explorarlos, grabarlos en dispositivos externos, así como eliminarlos e incluso modificarlos.[872]

En el centro de la neurotecnología se encuentran las interfaces cerebro computadoras (*Brain Computer Interfaces* "BCI", por sus siglas en inglés) que son dispositivos que conectan el cerebro humano con una computadora u otro dispositivo que se encuentra fuera del cuerpo humano. En este sentido, los dispositivos neurotecnológicos[873] pueden ser implantados o bien elementos no invasivos (gafas, casco, diadema, pulsera, etc) que a la vez pueden ser electrónicos o eléctricos, pueden ser chips, ópticos, magnéticos, acústicos, moleculares o químicos y utilizan cada vez más la IA.

Tal es así que observamos cómo los avances modernos de las neurociencias y las neurotecnologías han permitido desbloquear el cerebro humano y conocer aún más los procesos cerebrales y su relación con los estados mentales y el comportamiento observable.[874]

Es evidente que el desarrollo de las neurotecnologías puede generar impactos positivos en la calidad de vida y salud de las personas. En ese sentido, las neurotecnologías ofrecen un enorme potencial en el campo médico para el tratamiento de trastornos neurológicos

[872] Yuste Rafael, G. J. (2021). "It's time for neuro-rights". *Revista Horizon, N° 8*, p. 154 -156. & Rafael Yuste. (2017). Four ethical priorities for neurotechnologies and AI. Nature. Retrieved from <https://www.nature.com/articles/551159a>.

[873] Yuste, R. (2023). 1º Encontro Brasil-Chile de NEURODIREITO. *Perspectivas sobre a reforma constitucional brasileira para proteção jurídica da mente humana na era da inteligência artificial (AI) e da neurotecnologia.* Retrieved from <https://www.youtube.com/watch?v=9Yod5FcNwMo>.

[874] Ilenca, M. A. (2017). "Towards new human rights in the age of neuroscience and neurotechnology". Life Science and Policy, 13:5, p. 2.

y mentales. Hay más de trescientas mil aplicaciones móviles de salud diferentes disponibles en todo el mundo (un número que se ha duplicado en sólo cinco años), con un valor de mercado estimado que supera $100 mil millones.[875] En consecuencia, se pretende mejorar nuestra comprensión científica de la función cerebral humana y desbloquear los enigmas patológicos de varios trastornos neurológicos y mentales resistentes al tratamiento.[876]

Al mismo tiempo, este tipo de desarrollos se aplican cada vez más en contextos fuera del área de la salud, ingresando en campos como la enseñanza o el ámbito laboral y el del entretenimiento, entre algunos otros. A nivel mundial, el mercado de la neurotecnología está creciendo a un ritmo compuesto de tasa anual del 12 por ciento y se espera que alcance los $ 21 mil millones para 2026.[877, 878]

En ese sentido, existen dispositivos no invasivos que de manera muy simple pueden, por ejemplo, evaluar cómo se sienten los jugadores cuando son expuestos a distintos estímulos y en función de eso se pueden probar diferentes acciones, con el objetivo de potenciar el rendimiento. Asimismo, "también se está usando en el ámbito universitario para estudiar el desarrollo cognitivo en niños, en particular en poblaciones carenciadas", detallaron hace algún tiempo desde la empresa Emotiv.[879]

De este modo, se pueden tomar medidas para mejorar la atención, reducir el estrés o mejorar la concentración. Desde el sitio web de la empresa Emotiv[880] se puede acceder a comprar una serie de dispositivos disponibles, los que cualquier persona interesada puede

[875] "The Rise of mHealth Apps: A Market Snapshot," Best Practices (blog), Liquid State, March 26, 2018, updated November 12, 2019, The Rise of mHealth Apps: A Market Snapshot – Liquid State (liquid-state.com) Cit en The Battle for your Brain. Nita A. Farahany, 2023.

[876] UNESCO, C. I. (2021). Cuestiones éticas de la neurotecnología. UNESCO. Adoptado por el Comité Internacional de Bioética en su 28ª sesión de diciembre. Retrieved from <https://unesdoc.unesco.org/ark:/48223/pf0000378724>.

[877] Farahany, N. A. (2023). The Battle for your Brain. New York: St. Martin's Publishing Group.

[878] Expert Market Research, Global Neurotechnology Market Report and Forecast 2022–2027, Report Summary, accessed July 14, 2022, Neurotechnology Market Size, Share, Price, Growth, Forecast 2023-2028 (expertmarketresearch.com) Cit en The Battle for your Brain. Nita A. Farahany, 2023.

[879] Jaimovich, D. (2017, febrero 9). Así funciona Emotiv Insight, la vincha para mover objetos con la mente y controlar las emociones. INFOBAE. Retrieved from <https://www.infobae.com/tecno/2017/02/09/asi-funciona-emotiv-insight-la-vincha-para-mover-objetos-con-la-mente-y-controlar-las-emociones/>.

[880] <https://www.emotiv.com/>.

obtener desde los u$499 dólares. El dispositivo (vincha) recopila información y permite hacer un electroencefalograma (EEC) en apenas segundos y sin cables, se almacena y se puede utilizar para analizar el impacto de diferentes factores externos en la persona y sus emociones.[881]

Entendiendo la trascendencia que tiene el cerebro humano en sus funciones mentales y cognitivas, es que es necesario interpelarnos como sociedad respecto de hasta qué punto consideramos legítimas las interferencias en la actividad cerebral, si existen límites regulatorios actuales para su implementación o bien si necesitamos profundizar en regulaciones que precisen criterios claros para estos desarrollos.

## 22.2 Iniciativas regulatorias para minimizar el impacto en los derechos humanos

Durante el último tiempo se han desarrollado distintas iniciativas para minimizar el impacto de las neurotecnologías, en relación a la aplicación en IA. El enfoque siempre se proyecta en la protección de los derechos humanos, pero asimismo también hay cuestiones éticas en discusión.

Es decir, es importante resaltar que si bien son varias las tecnologías que presentan riesgos para los derechos humanos, nos referiremos específicamente a la IA con el uso de la neurotecnología.

En este sentido entendemos a la IA como la "constelación de procesos y tecnologías que permiten a las computadoras complementar o reemplazar tareas específicas que, de otro modo, serían desarrolladas por humanos, tales como, tomar una decisión o resolver un problema".[882]

Por otra parte, la IA podría utilizarse indebidamente y proporcionar herramientas de manipulación, explotación y control social.[883] Por lo que estos riesgos, podrían tener un mayor impacto en personas

[881] Jaimovich, D. (2017, febrero 9). Así funciona Emotiv Insight, la vincha para mover objetos con la mente y controlar las emociones. INFOBAE. Retrieved from <https://www.infobae.com/tecno/2017/02/09/asi-funciona-emotiv-insight-la-vincha-para-mover-objetos-con-la-mente-y-controlar-las-emociones/>.

[882] ONU, Reporte de la Alta Comisionada por Derechos Humanos, Michelle Bachelet (2021) "The right to privacy in the digital age", p. 2.

[883] Parlamento Europeo y Consejo Europeo (2021) "Reglamento por el que se establecen normas armonizadas en materia de IA (Ley de inteligencia artificial)", p. 24, par. 15.

que se encuentran en una situación especial de vulnerabilidad, tales como, niños, niñas y adolescentes, géneros y disidencias, personas con discapacidades, minorías étnicas y raciales, personas de la tercera edad, personas en situación de pobreza, entre otras.

A nivel internacional, varias organizaciones han comenzado a trabajar sobre el tema. Destacamos que en su "Recomendación sobre innovación responsable en neurotecnología", de 2019, la OCDE[884] menciona la necesidad de salvaguardas a la información mental.

Asimismo, en 2020, se publicó el borrador del Comité Internacional de Bioética de la Organización de las Naciones Unidas para la Educación, la Ciencia y la Cultura ("UNESCO") sobre *"Cuestiones Éticas de la Neurotecnología"*, que en sus conclusiones afirmó la necesidad de dar un marco al desarrollo de las mismas desde los derechos humanos, avanzando en algunas definiciones conceptuales.[885] En diciembre de 2021 se publicó el informe del Comité Internacional de Bioética de la UNESCO sobre cuestiones éticas y neurotecnología.[886] Finalmente, UNESCO convocó a una reunión que se desarrolló en el mes de julio de 2023 en su sede en París, en la cual se debatió la posibilidad de generar un marco de regulación global de las neurotecnologías, de forma similar a como estableció la Declaración Universal sobre el Genoma Humano en 1997.[887]

Por otro lado, fue importante el avance que se dio en octubre de 2022, cuando el Consejo de Derechos Humanos de Naciones Unidas aprobó por consenso la resolución A/HRC/51/L.3 sobre "Neurotecnologías y Derechos Humanos". Con ello, se inició un estudio sobre los impactos, oportunidades y desafíos de la neurotecnologías y se generó un proceso consultivo con actores estatales, multilaterales, sector privado y sociedad civil.[888]

[884] Disponible en <https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0457>.

[885] Disponible en UNESCO report draft.pdf – Google Drive Por otra parte, destacamos el impulso del Comité de Bioética del Consejo de Europa que publicó un Plan de acción estratégico sobre derechos humanos y tecnología en biomedicina. *Strategic Action Plan on Human Rights and Technologies in Biomedicine (2020-2025).* Consejo de Europa. Disponible en: <https://rm.coe.int/strategic-action-plan-final-e/1680a2c5d2>.

[886] UNESCO, C. I. (2021). Cuestiones éticas de la neurotecnología. UNESCO. Adoptado por el Comité Internacional de Bioética en su 28ª sesión de diciembre. Retrieved from <https://unesdoc.unesco.org/ark:/48223/pf0000378724>.

[887] Unesco (6 de junio 2023). "Es urgente establecer un marco ético sobre neurotecnología a escala internacional": <https://news.un.org/es/story/2023/06/1521747>.

[888] Entre los antecedentes que motivaron esta iniciativa, se mencionaron los avances a nivel Iberoamericano y la reforma constitucional chilena sobre protección a la actividad e información cerebral. Disponible en: A_HRC_51_L.3 (ohchr.org)>.

Asimismo, en febrero de 2020 la Comisión Europea publicó el "Libro Blanco sobre la IA: un enfoque europeo orientado a la excelencia y la confianza",[889] en el que se definieron las opciones que existen para compatibilizar los desarrollos en IA con la mitigación de los riesgos de ciertos usos de estas tecnologías. Esta propuesta convocó también el interés entre los actores vinculados a las neurotecnologías, por la profunda conexión que existen entre éstas y la IA. En especial, llamaron la atención sus contenidos sobre riesgos de discriminación y las referencias a las alteraciones de comportamiento que pueden inducir.

## 22.3 Principios Interamericanos en materia de neurociencias, neurotecnologías y derechos humanos:

A nivel regional interamericano se avanzó notablemente ya que contamos con el desarrollo de estándares internacionales con el trabajo elaborado por el Comité Jurídico Interamericano (CJI) de la OEA.

En este sentido, de forma novedosa en la materia el CJI aprobó la "Declaración sobre Neurociencia, Neurotecnologías y Derechos Humanos: Nuevos Desafíos Jurídicos para las Américas" (2021), que se constituyó en la primera de su tipo a nivel mundial,[890] en uno de sus apartados deja en claro la Declaración que no existen regulaciones específicas, por lo cual es indispensable realizar un llamado a los actores para que presten atención y queden a salvo los derechos humanos de las personas ante el vertiginoso desarrollo tecnológico.

En este sentido, la Declaración advierte que los avances de la neurociencia y el desarrollo de las neurotecnologías requieren de una reflexión profunda por parte de todos los sectores involucrados, hace un llamado a los Estados, al sector privado, la academia y el mundo científico y solicita la adopción de medidas concretas a cada uno de los actores que permitan que estas innovaciones contribuyan al bienestar común.

Luego, desde el Comité Jurídico Interamericano de la OEA continuaron los trabajos para desarrollar estándares más precisos

[889] Disponible en: <https://eur-lex.europa.eu/legal-content/ES/TXT/PDF/?uri=CELEX:52020DC0065>.

[890] CJI/DEC. 01 (XCIX-O/21). Disponible en: http://www.oas.org/es/sla/cji/docs/CJI-DEC_01_XCIX-O-21.pdf.

que contribuyan a orientar y armonizar las regulaciones nacionales necesarias en esta materia.

A partir de entonces se llevaron adelante distintas acciones,[891] entre los trabajos preparatorios, junto al Comité de Expertos que había colaborado en la redacción de la Declaración,[892] para profundizar la labor y generar principios mínimos para recomendar a los Estados el tratamiento de las neurotecnologías. Tal es así que el CJI aprobó un segundo informe de avance denominado "Proyecto de Principios Interamericanos en materia de Neurociencia, Neurotecnologías y Derechos Humanos" (2022).[893]

Llegando finalmente al 9 de marzo de 2023, cuando el CJI aprobó el documento sobre **Principios Interamericanos en materia de neurociencias, neurotecnologías y derechos humanos**.[894]

El desarrollo de estos Principios, es el resultado de un trabajo de análisis de las normas y estándares internacionales que ya se encuentran vigentes y son exigibles para los Estados. En este caso, se otorga de mayores precisiones sobre la materia específica del desarrollo de las neurotecnologías, para adelantarse y combatir cualquier situación que tienda a vulnerar los derechos humanos de las personas.

[891] El 21 de junio de 2022, se llevó adelante una audiencia pública ante la Comisión Interamericana de Derechos Humanos, a pedido de un grupo de expertos y expertas que tuvieron por meta principal la de identificar los desafíos, impactos, riesgos y eventuales vulneraciones a los Derechos Humanos que genera -o podría llegar a generar- el uso desregulado de las neurotecnologías, tanto de parte de los Estados como de los privados, en el ámbito de la salud de las personas. Asimismo, se pretendió mostrar y compartir recomendaciones para no incurrir en conductas riesgosas para los Derechos Humanos. La presente solicitud de audiencia fue efectuada por la Neuro Rights Initiative de la Universidad de Columbia, la Fundación Kamanau, la Red Pro Bono de las Américas, y Fundación Ronda, en conjunto con un grupo de expertos y expertas, de distintas nacionalidades y profesiones ligadas al mundo de la ciencia y del Derecho Internacional de los Derechos Humanos, quienes participaron en la elaboración de un documento para las Américas que diera cuenta de los desafíos e impactos de las neurotecnología en los Derechos Humanos y, en especial, en los neuroderechos. 184° periodo de sesiones de la CIDH: <https://www.youtube.com/watch?v=-JdUHdlXgdE>.

[892] El Comité de Expertos está integrado por: Eduardo Bertoni, Ciro Colombara, Francesca Fanucci, Verónica Hinestroza, Amelie Kim Cheang, Tomás Quadra Salcedo, Moisés Sánchez, Silvia Serrano Guzmán y Rafael Yuste.

[893] Proyecto de Principios Interamericanos en materia de Neurociencias, neurotecnologías y derechos humanos: <https://www.oas.org/es/sla/cji/docs/CJI-doc_673-22_rev1_ESP.pdf>. Este informe fue la base de un proceso de consulta para los Estados , además de una consulta pública abierta, donde se recibieron aportes que fueron siendo volcados al documento.

[894] Principios Interamericanos en materia de neurociencias, neurotecnologías y derechos humanos CJI/RES. 281 (CII-O/23) corr.1 <https://www.oas.org/es/sla/cji/docs/CJI-RES_281_CII-O-23_corr1_ESP.pdf>.

El documento establece diez apartados que refuerzan las garantías vigentes en protección de los derechos humanos en esta materia, teniendo como premisa fundamental la preservación de la identidad individual e integridad cognitiva frente a cualquier intervención neurotecnológica.

Además, determinados estándares derivan de una interpretación profunda sobre principios de protección de datos personales. Tal es así que establece que se protejan los derechos de las personas desde que se comiencen a diseñar las neurotecnologías hasta su despliegue definitivo, evaluación, comercialización y uso. Por otra parte, se busca brindar de mayor protección a los datos neuronales, enunciándolos como datos sensibles. Asimismo, establece la importancia de contar con consentimiento expreso para someterse a cualquier dispositivo que conlleve la manipulación de neurodatos.

Por otra parte, en relación a la protección del derecho a la igualdad se dispone el acceso equitativo a las neurotecnologías y garantizar la no discriminación de categorías que históricamente han sido objeto de discriminación: raza, color, género, nacionalidad, religión, condición social, entre otras. Sobre este mismo argumento se plantea la necesidad establecer límites claros y ejercer un control reforzado sobre el aumento de las capacidades cognitivas.

Por último, se promueven las estrategias para una eficiente gobernanza de las neurotecnologías, estableciendo organismos de supervisión y fiscalización y garantizando el acceso a una tutela efectiva.

Los Principios Interamericanos ofrecen una base de estándares sobre la que los Estados adopten medidas ante los avances del mundo científico.

Destacamos que hace unos días, en agosto de 2023, el CJI aprobó un nuevo temario para profundizar en este tema, se trata del abordaje en el impacto de las tecnologías basadas en IA en los derechos humanos, con especial enfoque en niños, niñas y adolescentes, tomando de base las neurotecnologías, las tecnologías inmersivas y emergentes con aplicación o basadas en IA.

## 22.4 Reciente sentencia de la Corte Suprema chilena sobre neurotecnologías

El 9 de agosto de 2023 la Corte Constitucional chilena se pronunció haciendo lugar a un planteo sobre la venta y comercialización en Chile del dispositivo Insight, de la empresa Emotiv, debido a que este dispositivo no protege adecuadamente la privacidad de la información cerebral de sus usuarios, infringiendo el derecho a la integridad mental, la integridad física y psíquica y el derecho a la vida privada.

Recordemos que el día 14 de octubre del año 2021, en Chile, se promulgó la Ley N° 21.383 que modificó la Constitución Política de la República, para establecer que el desarrollo científico y tecnológico estará al servicio de las personas y se llevará a cabo con respeto a la vida y a la integridad física y psíquica. Dispone la constitución que una La ley regulará los requisitos, condiciones y restricciones para su utilización en las personas, debiendo **resguardar especialmente la actividad cerebral, así como la información proveniente de ella** (art. 19, inc. 1°). En este sentido, entendió la Corte Constitucional que constituye un mandato directo de protección, además de diversos instrumentos internacionales que reconocen la relación entre ciencia y derechos humanos (C. S. Girardi c/ Emotiv Inc., 2023).

Ahora bien, analizaremos por separado algunas cuestiones que interesa destacar:

## 22.5 Dispositivo Emotiv Insight (neurotecnología)

De acuerdo al contexto científico que adelantamos anteriormente, la empresa de bioinformática Emotiv desarrolla productos de electroencefalografía portátil (en adelante, “EEG”),[895] incluidos neuroauriculares, kits de desarrollo de software (en adelante, “SDK”), softwares, aplicaciones móviles y productos de datos. Actualmente la empresa tiene su sede en San Francisco, Estados Unidos. Según la misma compañía, su misión es: “capacitar a las personas para

[895] La electroencefalografía (EGG) es el registro y evaluación de los potenciales eléctricos generados por el cerebro y obtenidos por medio de electrodos situados sobre la superficie del cuero cabelludo. Originalmente era utilizada en los campos de la psicología, medicina y neurociencia, pero actualmente es ampliamente utilizado en la interacción de seres humanos – computadora, gaming, neuromarketing, simulaciones y otros.

que comprendan su propio cerebro y acelerar la investigación del cerebro a nivel mundial".

Actualmente, Emotiv se destaca principalmente por el diseño de dos dispositivos: Emotiv Insight y EPOCx que son dispositivos de EEG que a través de técnicas no invasivas de neuroimagen de exploración funcional del sistema nervioso central, obtiene el registro de la actividad cerebral eléctrica de una persona en tiempo real. El EEG mide la actividad eléctrica del cerebro de manera muy simple, a través de la colocación de electrodos sobre la superficie del cuero cabelludo.

El dispositivo Insight, se posiciona como un dispositivo del tipo Interfaz cerebro Computador, no invasivo, conocido por sus siglas en inglés "BCI" Brain-Computer Interface (interfaz cerebro computador). Es inalámbrico y a través de una vicha cubre las ubicaciones frontal, temporal y parietooccipital alrededor del cerebro. Este dispositivo está diseñado para el uso diario utilizando sensores de polímeros hidrófilos, este dispositivo le permite al usuario leer sus emociones y desplazar elementos -tanto digitales como reales- con su mente.

Como lo expresa la misma empresa, a través de este dispositivo, se tiene acceso a datos de EEG, consistente en bioseñales eléctricas que incluyen información acerca de los gestos, movimientos, preferencias, tiempos de reacción y actividad cognitiva de su usuario.[896]

No obstante, si bien la tecnología actual todavía no permite leer los pensamientos, **las técnicas de neuroimagen tienen la capacidad de registrar la actividad cerebral. El sustrato mental de un individuo es producto de su actividad cerebral. La protección de ese fuero interno, de la experiencia subjetiva interna del sujeto, conforma una esfera individual única cuya protección se vincula de manera inseparable con la protección de su dignidad humana**.

## 22.6 Datos cerebrales (altamente sensibles) en la nube de la empresa

En este sentido y debido al interés del Sr. Guido Girardi Lavin en los dispositivos de neurotecnología y su preocupación con respecto a

[896] <https://id.emotivcloud.com/eoidc/privacy/privacy_policy/>.

los riesgos que pueden comprender a la privacidad de la información cerebral, con fecha 28 de febrero de 2022, compró el dispositivo Insight a través de la página web de Emotiv. Luego de pagar el envío a la empresa UPS, con fecha 21 de marzo de 2022 el dispositivo Insight llegó a su domicilio. Siguiendo las instrucciones del dispositivo y con el objeto de grabar y acceder a sus datos cerebrales, con fecha 7 de abril de 2022, creó una cuenta en la nube de datos Emotiv, llamada Emotiv Cloud. En dicha oportunidad, Emotiv solicitó aceptar los términos y condiciones de la empresa. Cuando el Sr. Girardi intentó empezar a grabar su información cerebral, Emotiv alertó que debido a que utilizó la licencia gratuita y no la licencia PRO, no podía exportar ni importar ningún registro de los datos cerebrales. En efecto, Emotiv señaló que dichos datos estarían retenidos en la nube de Emotiv hasta que éste comprara la licencia Pro. Minutos después, el sistema le alertó que la información cerebral grabada había sido exitosamente subida a la nube de Emotiv.

Consideró la Corte de Apelaciones que los neurodatos consisten en el conjunto de información relativa a la actividad cerebral obtenida mediante el empleo de neurotecnologías avanzadas. Los neurodatos forman parte del internet de los cuerpos, el avance de la IA nos sitúa en un horizonte muy próximo a la superinteligencia o IA de segundo nivel que ya no necesita intervención de los seres humanos.[897]

El punto aquí es saber **¿Emotiv protege adecuadamente la privacidad de la información cerebral de los usuarios, sobre todo en atención a la información altamente sensible como son los neurodatos?**

La respuesta de la empresa fue basarse en la seudonimización de los datos almacenados y por otra parte advirtió que ninguna medida de seguridad en 100% efectiva y que no pueden garantizar la seguridad de la información personal de los usuarios.

## 22.7 Consentimiento y finalidad (poco claros)

Se destaca también que otro de los argumentos de la demanda se centró en alegar que el uso del dispositivo y el almacenamiento

[897] No obstante ello y bajo el criterio de derecho de los usuarios a la autodeterminación informativa, se desestimó el planteo que luego llegó a la Corte Suprema. C.A. Girardi c/Emotiv Inc., 49852-2022 (Corte de Apelaciones 24 mayo, 2023).

de su información cerebral en la empresa Emotiv exponen a los usuarios al riesgo que sea compartida con terceros y que dichos datos sean objeto de investigaciones científicas e información estadística de libre uso.

Entonces, **¿es suficiente el consentimiento otorgado a través de la plataforma para que la empresa Emotiv almacene neurodatos y por otra parte que dentro de los términos y condiciones se indique que estos datos pueden ser compartidos con terceros para fines muy genéricos?**

La respuesta de la empresa fue manifestar que los usuarios deben prestar consentimiento expreso para el tratamiento de datos personales y cerebrales.

La Corte Suprema sobre esto referenció a la Ley chilena 20.120 que desarrolla un amplio articulo sobre la expresión del consentimiento ante la investigación científica en el ser humano (art. 11).

En este sentido, expresa la norma que existe consentimiento informado cuando la persona que debe prestarlo conoce los aspectos esenciales de la investigación, en especial su finalidad, beneficios y riesgos. Asimismo, aclara la Ley que debe haberse proporcionado información adecuada, suficiente y comprensible sobre ella. Por otro lado, deberá hacerse mención especial del derecho a no autorizar las investigaciones o de revocar el consentimiento en cualquier momento.

Por todo ello, consideró la Corte Suprema que "la explicación de la requerida, en orden a que los datos que obtiene de los usuarios de Insight, al ser anonimizados, pasan a ser información estadística de libre uso, omite como una cuestión previa la necesidad de contar con el consentimiento expreso de su uso para fines de investigación científica, distinta al registro estadístico, y expresamente regulada en Chile" (C. S. Girardi c/ Emotiv Inc., 2023).

Continuó la Corte Suprema que se descartaría la posibilidad que tal consentimiento pueda considerarse tácitamente prestado a través de otros consentimientos, por quienes en calidad de consumidores adquieren determinados aparatos y sobre el que deberían haberle requerido un consentimiento específico que indique además el propósito y fin de una determinada investigación con sus neurodatos.

## 22.8 El rol activo de Estado y el ¿principio precautorio?

Ingresando a la parte resolutoria del fallo, la Corte Suprema chilena consideró que ante el desarrollo de nuevas tecnologías que involucran cada vez más aspectos de la persona humana, que eran impensables hace algunos años atrás que pudieran conocerse, se debe otorgar especial atención y cuidado en su revisión por parte del Estado.

Es para destacar que este punto de **la sentencia recoge el principio precautorio traído del derecho ambiental, que se ejerce ante un potencial riesgo la acción del Estado a través de medidas de precaución.**

En este sentido, desde hace ya varios años atrás se entiende al principio precautorio como aquella actitud que deben observar quienes toman decisiones concernientes a una actividad que puede razonablemente ser riesgosa para la salud o seguridad de generaciones actuales o futuras.[898]

Esto, a los fines de prevenir y anticiparse a sus posibles efectos, además de proteger directamente la integridad humana, que incluye su privacidad, confidencialidad y los derechos propios de la integridad psíquica y del sujeto de experimentación científica (C. S. Girardi c/ Emotiv Inc., 2023).

Tal es así que consideró la Corte Suprema que en la actualidad sobre el uso de estas tecnologías "se hace absolutamente menester que previo a permitirse su comercialización y uso en el país, sean esta tecnología y dispositivos analizados por la autoridad pertinente, entendiendo que plantea problemáticas no antes estudiadas" (C. S. Girardi c/ Emotiv Inc., 2023).

Por lo anterior, es que expresó la Corte Suprema que la evaluación previa sobre "el manejo de datos que de él (dispositivo) se obtengan se ajuste estrictamente a la normativa aplicable" (C. S. Girardi c/ Emotiv Inc., 2023), a efectos de la comercialización y uso del dispositivo Insight, la debería hacer la Autoridad sanitaria , que en este caso es la autoridad de aplicación para autorizar productos o elementos de uso médico del Ministerio de Salud y la Autoridad

[898] Kamada, L. (2012). Del paradigma de la certeza al paradigma de la incertidumbre como criterio de decisión judicial en materia ambiental. SAIJ, Id SAIJ: DACF120104.

aduanera, para que evalúe otorgar el certificado de destinación aduanera correspondiente.

## 22.9 Privacidad mental (neuroderechos)

Además, para finalizar con la parte resolutoria del fallo, la Corte Suprema observó que las conductas que desarrolló la empresa Emotiv vulneraron las garantías constitucionales contenidas en art. 1 (integridad física-psíquica-actividad cerebral) y 4 (protección de datos personales) del artículo 19 de la Constitución.

Esto, **contemplando el alcance de la reforma constitucional que otorga de mayor resguardo y protección al desarrollo y uso de tecnologías que accedan y/o modifiquen la información cerebral sin respetar el derecho a la vida y a la integridad física y psíquica de las personas**. En este sentido, "se ordenó a eliminar toda la información que se hubiera almacenado en su nube o portales, en relación con el uso del dispositivo" (C. S. Girardi c/ Emotiv Inc., 2023) .

## 22.10 Necesidad de protección: ¿regulamos? ¿cómo?

Teniendo en cuenta los Principios Interamericanos de la OEA, las iniciativas que se vienen desarrollando a nivel internacional y nacional y la reciente jurisprudencia chilena, existen grandes desafíos por delante.

En este sentido, somos conscientes de la necesidad de establecer criterios regulatorios que pongan el claro los estándares internacionales sobre los nuevos avances de la ciencia y la tecnología, siempre con una perspectiva en base a los derechos humanos.

Para concluir y de manera muy breve, dejaremos algunos interrogantes que entendemos son claves para delinear regulaciones en términos de neurotecnologías, basadas en IA, o con aplicación en ella.

- ¿Es necesario fortalecer las leyes de protección de datos aclarando que los datos neuronales son datos personales sensibles? Que adquieran máxima protección en cesión, seguridad y transferencia
- ¿Estado activo que genere medidas previas de evaluaciones? ¿Traemos el principio precautorio a este tema?

- Es importante que se establezca el consentimiento expreso y específico y la finalidad esté detallada de manera adecuada ¿La seudonimización es suficiente para proteger la privacidad?
- ¿Es necesario implementar un modelo de cumplimiento denominado “accountability” o “responsabilidad proactiva” para la gobernanza de los riesgos?
- Autoridad de aplicación: ¿cuál es el modelo institucional ideal? ¿La sanitaria? ¿Es suficiente la autoridad de Datos Personales para asumir el control de la IA y neurotecnologías?

Es evidente que los neurodatos irrumpen el escenario jurídico planteando numerosos interrogantes. Conjugarán los criterios que establecen los recientes estándares regionales en materia de neurotecnologías sobre los límites en los cuales los Estados garantizarían de protección a los derechos humanos. Además, entendemos que se deben contemplar los argumentos valorados por el fallo de la Corte Suprema chilena analizados en el caso concreto. Todos los razonamientos deberían tomarse en consideración por las regulaciones que se desarrollen en el futuro.

En este sentido esta nueva forma de vivir (Farahany, 2023) está poniendo bajo una presión mucho más intensa la infraestructura regulatoria que se necesita para permitir y sostener que todos estos avances sucedan sin que se vulneren los derechos humanos. Es evidente que los desarrollos de las neurotecnologías y la IA exigen precisiones normativas de parte de los Estados, que –a su vez– necesitan ampararse en los Tratados Internacionales de Protección de los Derechos Humanos y las normas nacionales vigentes.

Tenemos grandes desafíos por delante para generar escenarios confiables en el desarrollo de estas nuevas tecnologías, sobre todo las tecnologías inmersivas que utilizan neurotecnología e IA.

Es urgente que se establezcan reglas claras y los derechos de las personas más vulnerables estén particularmente amparados, como son los niños, niñas y adolescentes, personas con discapacidad, mujeres y disidencias y personas mayores de edad.

The **authors** of this book are (in alphabetical order): Bhoomika Agarwal, Luca Belli, Bhavya Birla, Pedro Braga, Ian Brown, Jesús Javier Sánchez García, Walter Britto Gaspar, María Julia Giorgelli, Juan David Gutiérrez, Divij Joshi, Jonathan Mendoza Iserte, Natalia L. Monti, Thiago Moraes, Sizwe Snail ka Mtuze, Sarah Muñoz-Cadena, Melody Musoni, Ana Brian Nougrères, Nicola Palladino, Smriti Parsheera, Christian Perrone, Nadia Elsa Gervacio Rivera, Rama Vedashree Kamesh Shekhar, Jameela Sahiban, Attamongkol ("Atta") Tantratian, Rolf H. Weber, Wayne Wei Wang, and Yue Zhu.

This volume marks the beginning of activities of the Coalition on Data and Artificial Intelligence Governance (DAIG), a multistakeholder group established under the auspices of the United Nations Internet Governance Forum (IGF). The Coalition aims at fostering discussion of existing approaches to data and AI governance, promoting analysis of good and bad practices to identify what solutions should be replicated and which ones should be avoided by stakeholders to achieve a sustainable an effective data and AI governance.

To do so the DAIG Coalition aims at promoting collective studies and multistakeholder interactions to collect and discuss evidence, critically analyse existing regulatory and institutional arrangements, and propose policy updates in the areas of AI and data governance.

This first Annual Report of the Coalition aims at fostering reflections on transparency, accountability, and sovereignty in the context of AI governance, with a particular focus on experiences of Global South countries, to provide valuable contributions that could feed into IGF 2023 discussions. Particularly, this volume aims at answering pressing questions on the governance and regulation of AI systems, which are likely to have an enormous impact on the evolutions of our societies, economies, and democracies.

**This volume is the preliminary version of the 2023 Outcome Report of the UN IGF Coalition on Data and Artificial Intelligence Governance. It was presented at the United Nations Internet Governance Forum 2023, to receive feedback from the IGF community. The final version will be launched in 2024.**