Pre-print version of Belli L. Gaspar. W.B. The Quest for AI Sovereignty, Transparency and Accountability. Springer-Nature. (2025).

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 result of a participatory process developed by the Data and Artificial Intelligence Governance (DAIG) Coalition of the United Nations Internet Governance Forum (IGF). An early version of this volume's contribution was presented at the IGF 2023, in Kyoto, Japan, to receive feedback from the IGF participants. 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.**

Pre-print version of Belli L. Gaspar. W.B. The Quest for AI Sovereignty, Transparency and Accountability. Springer-Nature. (2025).

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***

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
Montevideo, October 2023

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 to foster 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 analyse 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**, which 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 analyses 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, analysing regulatory frameworks, and mapping harms and impacts are developed in 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, especially 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 system's 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) are 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 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.

TABLE OF CONTENTS

About the authors

Luca Belli, PhD is Professor of Digital Governance and Regulation at Fundação Getulio Vargas (FGV) Law School, Rio de Janeiro, where he directs the Center for Technology and Society (CTS-FGV) and the CyberBRICS project. Luca is also editor of the International Data Privacy Law (IDPL) Journal, published by Oxford University Press, and Director of the Computers Privacy and Data Protection conference Latin-America (CPDP LatAm). He is currently member of the Brazilian Presidency National Cybersecurity Committee, board member of the Global Digital Inclusion Partnership and member of the Steering Committee of the Forum for Information & Democracy. He is author of more than 50 publications on law and technology, exploring Internet access, data governance, cybersecurity, AI regulation, and digital transformation, 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 País, 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 at FGV in 2015. Master's 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). Currently, researcher in the CyberBRICS and the Data Regulations projects at FGV's Center for Technology and Society and Ph.D. candidate at the Public Policies, Strategies and Development Programme at the Economics Institute of the 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.

Liisa Janssens LLM MA is lead scientist of an interdisciplinary team of researchers at the department Military Operations, at the unit Defense, Safety & Security TNO (The Netherlands Organization for applied scientific research, founded by law in 1932 to enable business and government to apply knowledge). Liisa Janssens LLM MA combines theoretical and applied scientific research on the nexus of law, philosophy and AI technology (applied mathematics). In different engineering teams, she is the lead for formulating new questions on how to responsibly navigate AI development processes. In operational projects, for example, her role is to find new (technical) requirements and to build taxonomies and databases, all informed by ethics and the Rule of Law. This is necessary to establish (responsible) AI applications which can be implemented in society. She holds two positions as an expert in international governmental bodies, namely External Expert Member of NATO's Data AI Review Board (DARB), and she is appointed by the European Commission as a Reserve Member of the European Group on Ethics in Science and New Technologies (EGE).

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 the 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's 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's 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 has served as a graduate affiliate for the IU Center of Applied Cybersecurity Research. 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 bachelor's 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.

Jake Okechukwu Effoduh is an Assistant Professor at the Lincoln Alexander School of Law of Toronto Metropolitan University, Canada. He also served as the Chief Councillor of

Africa – Canada Artificial Intelligence and Data Innovation Consortium (ACADIC). ACADIC mobilizes AI and Big Data techniques in ethical ways to build resilient governance strategies for bilateral relations and increase societal preparedness for future global pandemics. Jake served as the human rights compliance expert, and ethics advisor for the consortium. Jake has been an international human rights lawyer for over 12 years with programmatic experience from working across Canada and 30 African countries. He holds two master's degrees in international law from the University of Oxford in the UK, and from York University in Canada.

Yves Poullet. Director of the CRID since its creation in 1979 until August 31, 2010, he supervised research in the field of new technologies, more particularly in matters of privacy, freedoms and the information society, the Internet and governance. In addition, he is Eméritus Professor at the University of Namur (1982-2017), Honorary rector (2010-2017), Co-chairman of the Namur Digital Institute (NADI), Associate Professor at the Catholic University of Lille (2017), Member of the Royal Academy of Belgium (2009), Member of the 'Chambre contentieuse' of the Belgian Data Protection Authority (2018), and Vice-chairman of the IFAP Programme in charge of the INFO-ETHICS Working Group (2020).

1. AI Sovereignty, Transparency and Accountability: 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 in AI governance are highly interconnected. Based on the previous work of Belli, we define AI sovereignty as the capacity to understand, develop, and regulate AI systems, and we argue that the assertion of AI sovereignty, transparency and accountability 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, this introductory chapter provides to the reader a useful guide to approach the concepts analysed in the different contributions collected in this volume, while highlighting the connections between such contributions. The chapter acknowledges that the concepts examined by the author of this book have been given much attention throughout policy debates and have been frame and defined in highly heterogeneous fashion, depending on the different public discourse arenas, and stakeholder interpretations. In this perspective, the goal of this chapter is to present the diverse issues that will be explored along the volume, enticing the reader to venture in the chapters of this work to further explore them.

1.1. How was this volume conceived and why? The first fruit of 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. This book was presented in its preliminary version at the inception meeting of the DAIG Coalition,[1] during the IGF 2023, in Kyoto Japan, to receive valuable feedback and critiques from participants. All comments have been incorporated in this final version and the authors would like to express immense gratitude to all IGF stakeholders having contributed to this effort.

1 See the IGF 2023 session organised by the DAIG Coalition to explore the issue "Can (generative) AI be compatible with Data Protection?" where the preliminary version of the book was presented for comments https://intgovforum.org/en/content/igf-2023-dc-daig-can-generative-ai-be-compatible-with-data-protection

Indeed, this participatory process embeds the *raison d'être* of the DAIG Coalition, which has been created with the aim of promoting studies and multistakeholder interactions to collect and discuss existing evidence and practices related to AI and data governance, critically analyse existing and proposed regulatory and institutional arrangements, and suggest policy updates in the areas of AI and data governance. Importantly, the DAIG Coalition 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. As an instance, many DAIG Coalition members form Latin American regularly meet at the CPDP LatAm conference[2] to debate what issues should be prioritised in the Coalition agenda and how to connect regional debates with global UN ones.

As the first Annual Report of the Coalition, this volume fosters reflections on transparency, accountability, and sovereignty in the context of AI governance, with a particular focus on experiences of Global South countries and provides valuable contributions that feed into IGF 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 evolution of our societies, economies, and democracies.

The fundamental assumption of this volume is to acknowledge that AI systems will have – and to some extent are already deploying – a transformative impact on how individuals, businesses, public administrations and state interact, therefore making it necessary to critically engage with the existing and proposed frameworks aimed at directing the development of such systems. In this perspective, we argue that 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. These concepts have been given much attention – and varying definitions in different contexts – throughout policy debates and have sparked increasing curiosity and questioning within many public discourse arenas. Hence, this volume should be seen as the beginning of a multistakeholder journey through the intricate landscape of AI sovereignty, transparency, and accountability. As the chapter of this book tellingly illustrate this journey may be rather turbulent, but a mounting number of works allows to illuminate the numerous facets of these key dimensions of AI governance.

Particularly, as governments and nations grapple with the complexities of asserting control over their AI technologies and data, it becomes increasingly clear that AI sovereignty is already or will become a key strategic priority. However, as the first part of this volume discusses, to achieve it, stakeholders will need to adopt a holistic approach that simultaneously addresses not only technical and regulatory aspects but also developmental, social, and geopolitical considerations. To address these challenges, multistakeholder cooperation is of essence. Offering a platform able to foster a multistakeholder cooperation

[2] The conference on Computers Privacy and Data Protection in Latin America (CPDP LatAm) takes place in Rio de Janeiro, every third week of July, and since 2023 included a meeting of the DAIG Coalition. See https://cpdp.lat/pt-br/programa/

including the perspectives of Global Majority in the discussion is the ultimate goal that underpins the creation of the DAIG Coalition, and this volume aims at demonstrating that such cooperation may produce concrete results, that can usefully inform public debates.

1.2. What do we mean by AI sovereignty, transparency and accountability?

The concept of AI Sovereignty was coined by Belli (2023) and defined as "a given country's capacity to understand, muster and develop AI systems, while retaining control, agency, and ultimately, self-determination over such systems" (2023, p.23). After being presented at Carnegie Digital Democracy Network Conference and at the UN IGF, respectively in May and October 2023, it rapidly became fashionable especially – and, to some extent, surprisingly – within US industry circles.

Notably, after the launch of the preliminary version of this volume at the UN Internet Governance Forum, in October 2023, multiple AI companies have jumped on the AI sovereignty bandwagon, coopting the terms and starting to brand their products as ideal solutions to become AI sovereign. Particularly telling examples have appeared, since early 2024, on the websites of NVIDIA and Oracle, two large AI equipment vendors, which published blogposts dedicated to "explaining" what AI Sovereignty is, basically arguing that governments merely need purchasing their products and services to become AI sovereign.[3]

It seems rather naïve to reduce AI sovereignty to the mere purchase of hardware and software from a foreign vendor. Even more naive if such purchase de facto creates a dependency form such vendor. In this sense, a sceptical observer may consider slightly self-interested and misleading the equation of technological dependence from one company as being AI sovereign.

As noted by Belli (2023) being AI sovereign requires recognising the essential role of research, development, governance and regulation with regards to at least considering Key AI Sovereignty Enablers (KASE), that are i) personal and non-personal data, ii) algorithmic models, iii) computational capacity, iv) meaningful connectivity, v) electric power, vi) capacity building, vii) cybersecurity and viii) AI-related risks. These elements are all equally important and interconnected and neglecting one or more may likely lead to create vulnerabilities in the entire AI Sovereignty Stack or failing to appropriately cope with existing vulnerabilities (Belli 2023).

In this context, this book starts form the assumption that the use of the KASE framework, exposed by Belli (2023), briefly presented below and elaborated further in the next chapter, has the potential to have a positive impact on how policymakers and other stakeholders design and implement AI strategies.

---

[3] Nvidia's blog states that "Sovereign AI refers to a nation's capabilities to produce artificial intelligence using its own infrastructure, data, workforce and business networks" adding that "[s]ince 2019, NVIDIA's AI Nations initiative has helped countries spanning every region of the globe to build sovereign AI capabilities." See Keith Strier. What Is Sovereign AI? NVIDIA's blog. (28 February 2024). https://blogs.nvidia.com/blog/what-is-sovereign-ai/ For Oracle, the very definition of Sovereign AI entails being a customer of its cloud computing services, as it succinctly states: "Achieve AI sovereignty with increased control over where you run your AI workloads and how you manage data and operate infrastructure with Oracle AI and Oracle Cloud Infrastructure (OCI) distributed cloud solution." See Oracle. Sovereign AI. Oracle's Cloud and Artificial Intelligence Blog. (s.d.) https://www.oracle.com/artificial-intelligence/sovereign-ai/

i) Personal and Non-Personal Data: Being able to construct solid data sovereignty, considered as "the capacity to understand how and why (personal) data are processed and by whom, develop data processing capabilities, and effectively regulate data processing, thus retaining self-determination and control" (Belli, Gaspar & Singh, 2024) is an essential component of AI sovereignty. Particularly, being able to have access to high-quality, diverse, and ethically sourced data is fundamental for training and improving AI models. Sovereignty over data allows nations to protect citizens' privacy, ensure data security, and regulate the use of sensitive information, thereby maintaining control over the development of AI systems, that reflect their cultural and societal norms, rather than being regulated by such systems.

ii) Algorithmic models: Software algorithms are the foundation of AI systems, enabling machines to perform tasks and make decisions. Importantly, algorithm development, deployment, and regulation are all equally important dimensions of algorithmic governance and algorithmic models can be both the subject and facilitator of regulation. By investing in algorithmic research and development (R&D), and crating sound policies and regulations, nations can develop cutting-edge algorithms that enable more accurate, efficient, and adaptable AI systems. This, in turn, empowers them to maintain a competitive edge in the global AI market, drive economic growth, and address pressing societal challenges. Furthermore, R&D on algorithmic models allows nations to tailor AI technologies to their specific needs and values, ensuring that AI systems align with their cultural, ethical, and legal frameworks. This is particularly important in domains such as healthcare, finance, and public policy, where the responsible and fair use of AI is paramount. By investing in R&D, nations can develop algorithms that are transparent, explainable, and accountable, thereby building public trust and facilitating the sustainable adoption of AI technologies.

iii) Computational Capacity: Training complex AI models and processing large datasets require substantial computational resources. Particularly, the most advanced AI systems, such as generative AI, can be remarkably computer-intensive due to their increased complexity. Ensuring the availability of sufficient computational capacity is critical for training and running complex AI systems. By investing in computational infrastructure, and regulating it in order to avoid concentration, nations can reduce dependence on external providers, accelerate AI research and development, and foster the growth of AI industries (Vipra & Myers West, 2023).

iv) Meaningful Connectivity: Reliable, well-performing, universally accessible internet infrastructure for an affordable price is essential for the deployment and operation of AI systems. 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 across sectors, contributing to the construction of a country's AI sovereignty.

v) Electric Power: AI systems require significant amounts of energy to operate, making access to reliable and sustainable power sources crucial. As AI systems grow in relevance and size, they require a stable and robust supply of electrical power to operate effectively. Ensuring the sustainability and reliability of power infrastructure is necessary for maintaining uninterrupted AI operations. By investing in renewable energy infrastructure, nations can support the growth of their AI industries, reduce carbon emissions, and ensure the long-term viability of AI applications.

vi) Capacity Building: Enhancing the digital literacy of the population through capacity building, training, and multigenerational education is essential not only to achieving a skilled AI workforce but also to fostering cybersecurity and, ultimately, national sovereignty. Developing a skilled workforce capable of designing, building, and deploying AI systems is particularly crucial. Hence, the objective of AI capacity building should be to cultivate and retain a talent pool that can drive innovation and competitiveness in the global AI market.

vii) Cybersecurity: The importance of cybersecurity cannot be overstated. AI systems are susceptible to cybersecurity threats and can be used to perpetrate cyber attacks, and AI critical infrastructure can come under attack. Indeed, AI can be leveraged by organisations to enhance their cyber defences, it can also be exploited by cybercriminals to launch targeted attacks at an unprecedented speed and scale, bypassing traditional detection measures (Belli, Gaspar et al. 2023). As AI systems become increasingly integrated into critical infrastructure, ensuring their security against cyber threats is paramount. Robust cybersecurity measures help protect AI assets, maintain public trust, and prevent the misuse of AI technologies for malicious purposes.

viii) AI-Related Risks: A comprehensive governance framework that addresses AI-related risks, encompassing ethical considerations, data protection laws, and risk-management tools is crucial for AI sovereignty. However, as we will stress, risk-based regulation is necessary but not sufficient to frame AI. Managing the ethical, legal, and societal implications of AI is essential for maintaining public support and ensuring the responsible development and deployment of AI systems. By establishing clear regulatory frameworks and, crucially, promoting transparency and accountability, nations can mitigate potential risks and harness the benefits of AI in a way that aligns with their fundamental values and priorities.

Importantly, the Key AI Sovereignty Enablers must be considered as interdependent, forming an AI Sovereignty Stack, which plays an instrumental role to determine a nation's ability to leverage AI for economic, social, and strategic advantage. By investing in these enablers, nations can reassert their sovereignty on AI technology, establishing a solid foundation for AI development, safeguarding their interests, shaping the evolution and adoption of AI in a way that promotes and preserves their values.

Furthermore, is important to emphasise that being AI sovereign does not mean being isolated into AI autarchy. On the contrary, it means being able to understand how AI system works and the impact they might have to be able to define one’s digital future, while cooperating with as many partners as possible, to increase trade and cooperation on fair terms and avoid being technologically dependent. This approach will allow us to discuss how nations in the Global South are striving to navigate the complexities of AI governance, to avoid becoming or continuing to be digital colonies4, forging innovative paths to harness the transformative potential of AI while safeguarding their interests and values.

4 The concept of digital colonialism is explored by an increasing number of authors. 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;

Intimately intertwined with the discourse on AI sovereignty are the imperatives of AI transparency and accountability. However, while these principles are considered as essential by a broad range of stakeholders, they often remain vaguely defined. From a governance and regulatory perspective, such lack of specification is unacceptable. As AI technologies increasingly permeate every aspect of society, transparency becomes essential to foster trust, enable informed decision-making, and mitigate potential risks and biases inherent in AI systems. However, transparency must be meaningful, (Belli et al., 2022) including specific substantial and formal criteria allowing stakeholder to operationalise such principle. In the luck of such meaningful transparency, it seems highly unlikely that individuals, societies, and governments may be able to understand how AI systems operate, assess their impacts, and hold stakeholders accountable for their actions.

Meaningful transparency is indeed instrumental to achieve accountability that, in turn, serves as a cornerstone of responsible AI governance, ensuring that AI systems are developed, deployed, and used in accordance with human rights, societal values, and national legislation. Accountability mechanisms, including oversight, auditability, and redress mechanisms, play a crucial role in promoting fairness, equity, and justice, especially as regards automated decision-making processes, thereby enhancing public confidence and legitimacy in AI systems.

These dimensions of AI governance entail some of the most thorny and consequential policy choices that our societies are called to make to shape the forthcoming future. The contributions presented in this volume try to shed light on the complexity of such choices and offer multistakeholder perspective on what potential solutions might be adopted. The chapters are organised around five thematic axes, briefly exposed in the following sections.

1.3. 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 in Brazil.

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; Benyera, E. (2021). The Fourth Industrial Revolution and the Recolonisation of Africa: The Coloniality of Data. Routledge.

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 regarding 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.

1.4. 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?", a question discussed in Rolf Weber's theoretical analysis of regulatory models in "Broadening the Horizon: New Concepts for AI Regulation". 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 adapt responsibility in the regulation of AI systems to the constituent parts of the AI supply chain. 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 in 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.

1.5. 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. Yves Poullet's chapter entitled "AI

and EU: A 'third way'?" provides a detailed account of the process by which the EU AI Act was discussed, the regulatory decisions contained therein and how they translate the European vision for AI in the broader context of EU policy. This birds-eye account of the act in context is then complemented by Nicola Palladino's "The Blind Watcher: Accountability mechanisms in the Artificial Intelligence Act", which 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 of how real-world institutional architectures can – or cannot – achieve their intended purpose.

In "Promoting the Transparency of AI-Generated Inferences", 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.

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 theatres.

1.6. 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 feasibility 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. These considerations are made evident in the analysis of the DigiYatra project in Indian airports, where issues in all layers are identified regarding transparency of processes.

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.

Michael Karanicolas' chapter on "Developing AI Standards that Serve the Majority World" argues that institutional and regulatory architectures tailored to "Global Minority" countries' needs, although frequently adopted in "Global Majority" through diffusion, are often sub-optimal outside their originating contexts. Thus, there is a need to develop AI standards beyond the "auspices of a handful of powerful regulatory blocs", calling for the inclusion of the Majority World into standard-setting processes in international fora.

Lastly, Jake Okechukwu Effoduh's chapter entitled "(Re)Examining the Concept of Regulation in AI Governance" does a deep dive into the challenges of regulating AI due to its inherent characteristics, such as its deterritorialized and fast moving nature. With examples from African countries, the chapter highlights the need to develop regulatory pathways that adapt to particular countries' needs while also taking advantage from the international discussions already in place.

1.7. 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. Braga and Perrone's "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)" 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.

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.

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 remaining 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 in "International efforts aimed at promoting AI transparency and/or accountability". 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, in "Regulatory Aspects of AI in Argentina".

Natalia Monti's chapter, entitled "AI and neurotechnologies: the need for protection in the face of new crossroads", 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.

1.8. 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. This volume provides theoretical and practical grounds to counter that vagueness, giving substance to the concepts commonly surrounding discussions of AI regulation. Thus, it draws a path from detailed discussions of the concepts of transparency and accountability to the proposal of concrete measures in the form of frameworks, mechanisms or processes. As a reflection of the multistakeholder nature of the roster of authors, a multitude of perspectives are provided, concerned with the implementation of AI in public and private settings.

The volume also looks at AI regulation as a matter of public policy and ensuring rights. This is present in discussions such as the key AI sovereignty enablers framework, the need for strict trade secret regulations that leave space for the rights of subjects affected by AI tools, and the need to evaluate military applications under a strict lens of legality and responsible use. These are discussions that do not shy away from hard questions concerning decisions in implementing AI systems that affect collective and individual rights, not only through their immediate effect, but also via long-reaching consequences of policy decisions.

Finally, this volume is also dedicated to discussing AI policy and regulation from the perspective of diverse nations and blocs. This reiterates that markers such as "majority" and "minority" world, global "north" and "south" or "centre" and "periphery" nations are designed to encompass and highlight structural imbalances in the international system which call for context-specific policy and regulation, including when dealing with AI. 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.

References

AVILA PINTO, R. Digital sovereignty or digital colonialism? New tensions of privacy, security and national policies. *SUR: International Journal on Human Rights*, v. 15, n. 27, 2018.

Belli L., Gaspar, W.B., Singh Jaswant, S. Data sovereignty and data transfers as fundamental elements of digital transformation: Lessons from the BRICS countries. Computer Law & Security Review. Special issue on Digital Transformation in the BRICS Countries. Volume 54. (2024). https://doi.org/10.1016/j.clsr.2024.106017

Belli, L. "To Get Its AI Foothold, Brazil Needs to Apply the Key AI Sovereignty Enablers (KASE)." In Steven Feldstein (Ed.) New Digital Dilemmas: Resisting Autocrats, Navigating

Geopolitics, Confronting Platforms. Washington, DC: Carnegie Endowment for International Peace. (2023). http://dx.doi.org/10.2139/ssrn.4465501

BELLI, L. et al. Towards Meaningful and Interoperable Transparency for Digital Platforms 2022 Outcome of the UN IGF Coalition on Platform Responsibility. UN IGF, 2022. Available at: https://www.intgovforum.org/en/filedepot_download/57/23886.

Belli, L., Gaspar W.B., et al. Cibersegurança: uma visão sistêmica rumo a uma proposta de Marco Regulatório para um Brasil digitalmente soberano. FGV Direito Rio. (2023). https://repositorio.fgv.br/items/4814c750-6b42-4d48-b8bb-302ce467b4ea

BENYERA, E. The Fourth Industrial Revolution and the Recolonisation of Africa: The Coloniality of Data. Routledge, 2021.

COULDRY, N.; MEJIAS, U. The costs of connection: How data Is colonizing human life and appropriating it for capitalism. Stanford, CA: Stanford University Press, 2019.

Vipra, J. & Myers West, S. Computational Power and AI. AI Now Institute. (2023) https://ainowinstitute.org/publication/policy/compute-and-ai

PART 1:
FRAMING THE AI SOVEREIGNTY DEBATE

Pre-print version of Belli L. Gaspar. W.B. The Quest for AI Sovereignty, Transparency and Accountability. Springer-Nature. (2025).

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), primarily through a risk-based approach. The main goal of this paper is to emphasise that the regulation of AI risks is only one of the essential elements that need to be considered to achieve AI Sovereignty. Importantly, the paper defines AI Sovereignty as the capacity of a given country to understand, develop and regulate AI systems, thus 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 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 framework to Brazil, to investigate whether the existing 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. Lastly, it argues that national governments should strive to revert AI dependency and build AI Sovereignty, to avoid a scenario of digital colonialism, while keeping on promoting cooperation and striving to eschew protectionism.[1]

Introduction

As a transformational technology (Jarvenpaa & Ives, 1996), Artificial Intelligence (AI) will have a global impact and considerable ramifications for national economies, democracies, and societies. While many countries are developing regulatory proposals aimed at framing AI risks (Belli, Curzi & Gaspar, 2023), the main goal of this paper is to emphasise that the regulation of AI-related risks is only one of the essential elements that need to be considered to achieve AI Sovereignty.

AI Sovereignty is a new and not universally defined concept. In this paper, I put forward a definition of this concept, building on previous work on digital sovereignty (Jiang

[1] This paper is an extended version of an essay presented at the Digital Democracy Network Conference, in May 2023, organised by the Carnegie Endowment for International Peace, and published in the edited collection resulting from the conference. See Belli, L. To Get Its AI Foothold, Brazil Needs to Apply the Key AI Sovereignty Enablers (KASE), in Feldstein S. (Ed.) (2023). Return to New Digital Dilemmas: Resisting Autocrats, Navigating Geopolitics, Confronting Platforms. Carnegie Endowment for International Peace. The author would like to thank Steven Feldstein, the Carnegie team, 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.

& Belli 2024) and, particularly, what I have previously described as "Good Digital Sovereignty," (Belli, 2023) thus considering AI Sovereignty as the capacity of a given country to understand, develop and regulate AI systems. While acknowledging that the concept of digital or cyber sovereignty may have some controversial connotations, flirting with authoritarianism and protectionism, I argue that AI Sovereignty should be seen in its positive conception of essential enabler of agency and self-determination[2] over AI systems.

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 being ineffective to mitigate risks and even counterproductive, leading to increased technological dependency.

I argue that sound governance[3], regulation, research, and development in all the elements of the AI value chain are essential to achieve economic growth, social justice, and industrial leadership, which are key pillars of (AI) sovereignty. Indeed, by avoiding the adoption and implementation of exclusively foreign AI systems, governments can endeavour to eschew the transformation of their countries into digital colonies, whose dependence on foreign AI can hardly be reversed. Importantly, the purpose of this paper is not to advocate for AI autarchy and isolationism, nor to deny the ample range of benefits that digital trade and cooperation can produce. On the contrary the goal of this author is to discuss how countries could achieve a sufficient level of strategic autonomy, diversifying their AI value

---

[2] 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 (2017: 35-64) ; Belli *et al.* (2023: 69-94).

[3] 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. See Belli (2016: 17-132).

chains, and being able to grasp the functioning of AI systems, developing such systems, rather than being mere consumers dependent from foreign suppliers, and regulating affectively AI according to their national values.

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 allow countries to approach AI from a more strategic perspective, thus reducing their exposure to the technological choices of foreign (private or public) actors, and simultaneously increasing their agency and self-determination over and through AI systems.

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. The stack should be organised through a dedicated AI Sovereignty Governance System, allowing representatives of authorities in charge of overseeing each KASE to cooperate with representatives of authorities from related sectors – including with regulators of transversal sectors such as competition, consumer protection, data privacy, financial services, energy, and telecom infrastructure, as well as financial and developmental institutions – to facilitate smooth organisation, cooperation 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 an urgent policy priority, to avoid that AI dependency becomes the norm. Indeed, any type of dependence, once it is established, it becomes extremely difficult to reverse. The KASE framework discussed in the next section require considerable planning, resources, and implementation capacity, but should be seen as a highly strategic objective 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.

2.1. 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, emphasising the relevance and the interconnectedness of the Key AI Sovereignty Enablers (KASE). These elements are instrumental for ensuring that a country can understand, develop, and regulate 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[4]) entities' exercise of AI Sovereignty.

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

[4] For a better understanding of how corporate actors can deploy a private form of soverignty, see Belli (2022)

considerable research, especially concerning its interaction with data governance[5], includes a wide range of critical sectors such as defence, national security, infrastructural management, healthcare, and justice.

As mentioned previously, it seems important to emphasise that the capacity to understand the functioning and impact of AI systems, being able 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 aimed at regulating risks. On the contrary, the achievement of AI Sovereignty entails the capacity to exercise agency and self-determination on at least eight different dimensions 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 research, development, governance and regulation of (personal) data, algorithmic models, computational capacity, meaningful connectivity, reliable electrical power, digital literacy, cybersecurity, and last, but not least, an appropriate framework regulating AI risks. The next section analyses them, in the context of Brazil.

2.2. 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.

2.2.1. Data

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 (open data), 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 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 comprehensive data policies will indubitably have a competitive advantage, constructing their AI (and data) sovereignty. However, it is important to emphasise that very 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 governance frameworks at regional level or within existing intergovernmental fora,[6] so that national data assets can be processed and transferred under substantially equal norms.

---

[5] A selection of telling examples is provided in the publication section of the CPDP LatAm conference website available at https://cpdp.lat/en/publications/

[6] 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

Importantly, such frameworks should go beyond personal data and adopt a more comprehensive approach to data governance, including norms defining the usage and reusage of open data and copyrighted information as well as guarantees against misuse of sensitive and confidential information. This strategy would allow to mitigate risks and reap the benefits 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 (De Magalhães Santos, L. G., & Dhaou, 2022) and by adopting a last-generation data protection framework, the *Lei Geral de Proteção de Dados* or LGPD[7]. The enforcement of the LGPD, however, remains still very timid and at an embryonic stage, especially as regards (generative) AI systems -Belli, 2023c).

Furthermore, personal data collection is considerably concentrated in the hands of a few foreign tech giants, primarily because of so-called zero-rating mobile Internet plans[8] subsidising access to social media, 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 (Belli, 2021) in the lack of a general Cybersecurity law and given the lack of regulation on personal data security.

2.3. Algorithmic models

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. Such risks should be considered carefully, especially considering the fact they can vary enormously depending on the AI systems at stake – e.g. foundational models present very different risks compared with algorithms trained on hyper personalised and localised data bases. 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

---

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 Belli et al. (2024).

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

[8] See http://www.zerorating.info/.

owning proprietary software provides a considerable competitive advantage and allows for embedding normative values according to national specificities. Investing in research and development of algorithmic tools, 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, 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 (Belli, 2023d; Jiang & Belli 2024), 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 adopting foreign software. It is important to stress that a revival of national support of open source could be a meaningful way to strengthen domestic AI development, especially given the existence of several interesting options of open-source models, such as GPTNeo and BLOOM and, to minor extent, Falcon, Gemma or Llama, on which novel AI systems can be built.[9]

Despite political turbulence, over the past two decades, Brazil has developed several industrial policy instruments aimed at fostering the national software industry and is currently planning the adoption of a new industrial policy which will include digital transformation as one of its core pillars.[10] However, the software development sector has not become as thriving as it could, primarily due to a lack of consistency in software-related policies over the past decades 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[11], or by organising capacity building efforts aimed at fostering demand, as South Korea did in the late 1990s.

2.4. Computational Capacity

[9] The varying degree to which these models can be considered as "open" is the object of intense academic and policy debates. For an interesting overview of the issue, see Widder, West and Whittaker (2023).

[10] The Brazilian Ministry of Development, Industry, Commerce and Services announced in January 2024 its new industrial policy, based on six core missions. Mission number 4 aims at "digitally transforming 90% of all Brazilian industrial companies (today only 23.5% is digitalized) and tripling the share of national production in the new technology segments." See Ministério do Desenvolvimento, Indústria, Comércio e Serviços. (2024).

[11] See https://indiastack.org/.

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 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.[12]

More recently, one of the key priorities identified by the new Brazilian industrial policy is the support of the national semiconductor industry[13] as well as the more wholistic promotion of a Brazilian AI Plan (PBIA in its Portuguese acronym), titled "AI for the Good of All". The plan includes around 4 billion USD in investments primarily dedicated to the increase of the national computational capacity, the development of domestic algorithmic models and the promotion of educational initiatives. (Brazil, 2024)

The Brazilian strategic posture towards computational capacity and its component is still embryonic but is interesting to note that the increasing understanding of the relevance that industrial policy plays regarding this issue is far from being a Brazilian peculiarity. Indeed, according to data from Chinese semiconductor industry research firm JW Consulting, the Chinese government has allocated more than CNY 2.1 trillion (US$290.8 billion) to semiconductor-related investment, between 2021 and 2022, supporting 742 investment projects in 25 Chinese provinces and regions (Lin, 2023). The issue has been the object of gargantuan public investments even in the United States, which have traditionally play down the role of industrial policy, through the CHIPS and Science Act, 2022, which established a funding equivalent to roughly $280 billion to turbocharge domestic research and development of semiconductors.

Moreover, it is essential to emphasise that while computing power is key (Vipra & West, 2023), 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 provided by two major US tech companies in Brazil, which

[12] 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

[13] The overview of the new industrial policy missions and priorities identified by the Brazilian Government in January 2024 is available at https://www.gov.br/mdic/pt-br/composicao/se/cndi/arquivos/missoes-politica-industrial.pdf It is important to stress, however, that at the time of this writing the specific budget and the detailed planification of how such budget will be spent had not been released yet, thus not allowing the author to assess the soundness of the announced commitments.

are supplied nationally and do not even mention how they comply with the Brazilian LGPD, despite the law being fully in force since 2021 (Chacon, Bawden & Xavier Morales, 2022).

Lastly, it is important to stress that the promotion of open-source AI can have an important role to play in the mitigation of compute concentration, as it allows the development of cheaper and more distributed AI systems reducing the need to reply on highly concentrated compute resources. Indeed, the global cloud computing market relies on a handful of corporations, with the "Big Three" - i.e. Amazon Web Services, Microsoft Azure and Google Cloud – currently accounting for two thirds of the growing cloud market, with a remarkable 20% year-on-year growth rate at the end of 2023, largely due to "generative AI technology and services [that] have had a major impact, helping to further boost cloud spending" (Synergy Research Group, 2024).

2.5. 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 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 (TIC domicílios, s.d.). 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 (NIC.br, 2024).

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 (IDEC 2021). Such plans are based on defining a limited monthly data volume for users and not counting the data consumption of a few applications selected by the mobile internet operators from the existing data cap. As such user attention ad user data collection is concentrated in a remarkably limited number of services, perceived as free by users, but whose access is de facto paid with personal data rather than with money.

The fact that zero rated apps typically are dominant social media platforms, makes it particularly challenging for any other business to compete, being almost impossible to develop personal data sets which are as complete as those owned by the established zero-rated platforms and can be used to train AI models, or event more difficult to achieve a sufficiently large user base that can be employed to test and refine new AI systems.

2.6. Reliable electrical power

As AI systems grow in relevance and size, they require a stable and increasingly relevant supply of electrical power (Luccioni, 2023) to operate effectively. Ensuring reliable

power infrastructure and access to affordable and sustainable electricity is necessary for maintaining uninterrupted AI operations, while balancing the electrical needs of AI development with the needs of the local stakeholders. 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 (Ministério de Minas e Energia, 2023).

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 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.

Furthermore, it is important to emphasise that abundance of clean energy should be confused with ubiquity of such energy. In practice this means that it would be relatively easy to power multiple data centers in the proximity of hydropower or wind-power sources, but it is highly impractical and likely unsuccessful to plan the development of large computational facilities distant from such sources. In this perspective, it hard to understand the rationale behind the Brazilian government plan to enormously expand the Santos Dumond supercomputer located in Petropolis, a minor city in the Rio de Janeiro State, far from both energy sources and the actors that could make use of the computing infrastructure it provides.

Lastly, an intimately intertwined element is the presence of large water supply to cool AI computing infrastructure. Indeed, while AI companies are particularly opaque about their energy and water consumption, recent research illustrates that the water footprint of some of the largest AI models is clearly unsustainable (Pengfei Li et al., 2023). Indeed, only for "training GPT-3 in its data centers, Microsoft was estimated to have used 700,000 liters — or about 185,000 gallons — of fresh water. That's enough water to fill a nuclear reactor's cooling tower" (Gendron, 2023).

2.7. 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 (Belli et al. 2023). 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[14].

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 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.

2.8. Strong cybersecurity

As noted in (Belli, 2024), AI has transformed the cybersecurity landscape over the past decade, leading to an increase in the frequency, impact, and sophistication of cyberattacks. While AI can be leveraged by organisations to enhance their cyber defences, detecting cyberthreats and improving decisions about how to react, it can also be exploited by cybercriminals to launch targeted attacks at an unprecedented speed and scale, bypassing traditional detection measures.

Indeed, the increasing use of AI systems in a wide range of processes in various safety-critical sectors – such as health, justice, autonomous vehicle-management, etc. – creates numerous new, and sometimes unpredictable, risks and can open new avenues in attack methods and techniques. Such risks may be maximised when AI is deployed for automated decision making, directly affecting both individuals and organisations, thus leading legislators around the world, including in Brazil, to consider appropriate risk regulations aimed at framing AI systems

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 (Belli et al. 2023), 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 (Brazil, 2022), it must be noted that this positive advancement must be considered again with a grain of salt.

Indeed, despite adoption of multiple sector-specific regulations, 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 (Belli et al. 2023) and by a Draft Bill formulated by the Brazilian Presidency (PNCiber Draft Bill, n.d.). 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,

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

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

Moreover, despite the adoption of new sectoral regulation and legislation that can be used to improve cybersecurity, such norms are frequently poorly implemented. A telling example is offered by the "security by design" obligation introduced by art 46 of the Brazilian data protection law, LGPD. In the first four years since the entry in force of LGPD and the establishment of the national data protection regulator, the ANDP, not a single entity has been sanctioned for lack of compliance with this norm, despite the exponential increase of information security incidents – most of which involve personal data – and the

2.9. Appropriate regulatory framework

A comprehensive governance framework that encompasses ethical considerations, data protection laws, and a risk-based approach to AI regulation is crucial for AI sovereignty. Establishing clear guidelines and standards for AI development, deployment, and usage is instrumental to ensure responsible and accountable AI practices. In this perspective, the Brazilian Congress is elaborating a new AI Regulatory Framework (Brazil, 2023) 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 and needed, even if still ongoing, it is not yet clear to what extent it will be able to effectively steer the evolution of AI in the country. One of the most discussed versions of the proposed framework, enshrined into Bill 2338/2024, includes many normative provisions, inspire from the EU AI Act, which aim at regulating key issues such as AI systems transparency, data security, data governance or risk management. Adopting flexible norms is a welcome regulatory strategy to craft laws that are future-prof and can adapt to technological evolution. However, to be meaningful the flexible clauses must also be matched with an implementation mechanism that allows their specification through regulation or standardisation.

In the absence of such specifications, the law risks highly ineffective and vague, rather than flexible, and become simply impossible to implement. 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. Crafting a flexible data protection law has been an essential step based on the national consensus regarding the need for a Brazilian data protection framework. But the effectiveness of the regulatory strategy risks being considerably jeopardised when the pressing task of specifying the law is either undefined or, as it happens in the Brazilian context, is attributed to a regulator that is so under-resourced to seem purposefully created to be "ineffective by design" (Belli, 2022).

2.10. 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 focusing only on risk regulation and ignoring the utmost importance of all the other fundamental elements that compose the KASE framework. 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 integrated approach seems to be absent from the current Brazilian "strategic" vision for AI. Indeed, anyone analysing the 2021 Brazilian Artificial Intelligence Strategy (EBIA) (Gaspar, 2022) 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. It is, therefore, enormously heartwarming to read that the Brazilian Ministry for Science and Technology has decided to reformulate the EBIA, signalling to have realised the lack of vision, objectives and, ultimately, feasibility of the previous version (MCTI, 2023).

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, strengthening and diversifying supply chains, increasing the digital literacy of the population, fostering strategic investments and partnerships, and safeguarding the security of critical AI infrastructure, besides regulating AI risks.

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 to build an AI Sovereignty Stack. Such an effort might be especially herculean 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[15] is likely to exacerbate.

References

[15] See e.g. Avila Pinto, R. (2018); Couldry, N. & Mejias, U. (2019).

AVILA PINTO, R. Digital sovereignty or digital colonialism? New tensions of privacy, security and national policies. SUR: International Journal on Human Rights, v. 15, n. 27, 2018.

BELLI, L. Brasil precisa reconstruir sua soberania digital. Estadão, 1 mar. 2023. Available at: <https://www.estadao.com.br/politica/blog-do-fausto-macedo/brasil-precisa-reconstruir-sua-soberania-digital/>.

BELLI, L. Building Good Digital Sovereignty through Digital Public Infrastructures and Digital Commons in India and Brazil. G20's Think20 (T20), jun. 2023. Available at: <https://t20ind.org/research/building-good-digital-sovereignty-through-digital-public-infrastructures/>.

BELLI, L. De la gouvernance à la régulation de l'Internet. Paris: Berger-Levrault, 2016, p. 17-132.

BELLI, L. et al. Cibersegurança: uma visão sistêmica rumo a uma Proposta de Marco Regulatório para um Brasil Digitalmente soberano. CyberBRICS, 2023. Available at: <https://cyberbrics.info/ciberseguranca-uma-visao-sistemica-rumo-a-uma-proposta-de-marco-regulatorio-para-um-brasil-digitalmente-soberano/>.

BELLI, L. et al. Cibersegurança: uma visão sistêmica rumo a uma proposta de marco regulatório para um Brasil digitalmente soberano. FGV Direito Rio, 2023, p. 69-94. Available at: <https://bibliotecadigital.fgv.br/dspace/handle/10438/33784>.

BELLI, L. Network Self-Determination and the Positive Externalities of Community Networks. In: BELLI, L. (Ed.). Community Networks: The Internet by the People for the People: Official Outcome of the UN IGF Dynamic Coalition on Community Connectivity. FGV, 2017, p. 35-64. Available at: <https://www.intgovforum.org/en/filedepot_download/4391/1132>.

BELLI, L. New Data Architectures in Brazil, China, and India: From Copycats to Innovators, towards a post-Western Model of Data Governance. IJLT, n.d. Available at: <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>.

BELLI, L. Structural Power as a Critical Element of Digital Platforms' Private Sovereignty. In: CELESTE, E.; HELDT, A.; KELLER, C. I. (Eds.). Constitutionalising Social Media. Hart, 2022. Available at: <https://lucabelli.net/2021/08/10/structural-power-as-a-critical-element-of-social-media-platforms-private-sovereignty/>.

BELLI, L. The largest personal data leakage in Brazilian history. openDemocracy, 2021. Available at: <https://www.opendemocracy.net/en/largest-personal-data-leakage-brazilian-history/>.

BELLI, L. To Get Its AI Foothold, Brazil Needs to Apply the Key AI Sovereignty Enablers (KASE). In: FELDSTEIN, S. (Ed.). Return to New Digital Dilemmas: Resisting Autocrats, Navigating Geopolitics, Confronting Platforms. Carnegie Endowment for International Peace, 2023.

BELLI, L. Why ChatGPT does not comply with the Brazilian Data Protection Law and why I petitioned the Regulator. MediaNama, 20 jul. 2023. Available at: <https://www.medianama.com/2023/05/223-chatgpt-brazilian-data-protection-law-ai-regulation/>. BELLI, L.; GASPAR, W.; COUTO, N. Por que o ChatGPT descumpre a LGPD – parte 2. Jota, 25 ago. 2023.

BELLI, L.; CURZI, Y.; GASPAR, W. B. AI regulation in Brazil: Advancements, flows and need to learn from the data protection experience. Computer Law & Security Review, v. 48, 105767, 2023. Available at: <https://www.sciencedirect.com/science/article/pii/S0267364922001108>.

BELLI, L.; JIANG, M. (Eds.). Digital Sovereignty from the BRICS Countries. Cambridge University Press, forthcoming.

BELLI, L.; NOUGRÈRES, A. B.; MENDOZA ISERTE, J.; PALAZZI, P. A.; ANGARITA, N. R. Hacia un modelo latinoamericano de adecuación para la transferencia internacional de datos personales. Centro de Tecnologia y Sociedad de Universidad de San Andrés, 2023.

BRAZIL. (2 June 2024). IA para o bem de todos: Proposta de Plano Brasileiro de Inteligência Artificial 2024-2028. CNCT. https://www.gov.br/mcti/pt-br/acompanhe-o-mcti/noticias/2024/07/plano-brasileiro-de-ia-tera-supercomputador-e-investimento-de-r-23-bilhoes-em-quatro-anos/ia_para_o_bem_de_todos.pdf/view

BRAZIL. Lei n. 14.533, de 11 de janeiro de 2023. Available at: <https://www.planalto.gov.br/ccivil_03/_ato2023-2026/2023/lei/L14533.htm>.

BRAZIL. PL 2338/2023 - Senado Federal. Available at: <https://www25.senado.leg.br/web/atividade/materias/-/materia/157233>.

BRAZIL. Brazil rises in international cybersecurity ranking. Serviços E Informações Do Brasil, 24 jun. 2022. Available at: <https://www.gov.br/en/government-of-brazil/latest-news/2022/brazil-rises-in-international-cybersecurity-ranking>.

Chacon, G., Bawden, H, Xavier Morales, L. Termos De Uso e Políticas De Privacidade do Google Workspace for Education e Microsoft 365 (Office 365 Educação). LAPIN. (2022). https://zenodo.org/records/7718863

COULDRY, N.; MEJIAS, U. The costs of connection: How data Is colonizing human life and appropriating it for capitalism. Stanford, CA: Stanford University Press, 2019.

CPDP LatAm. Publications - CPDP LaTAM 2023. CPDP LatAm 2023, 18 jul. 2023. Available at: <https://cpdp.lat/en/publications/>.

CYBERBRICS. Cybersecurity and digital sovereignty: a new path for Brazil. CyberBRICS, 24 fev. 2023. Available at: <https://cyberbrics.info/cybersecurity-and-digital-sovereignty-a-new-path-for-brazil/>.

DE MAGALHÃES SANTOS, L. G.; DHAOU, S. B. Open Data and Emerging Technologies: Connecting SDG Performance and Digital Transformation. Available at: <https://cyberbrics.info/open-data-and-emerging-technologies-connecting-sdg-performance-and-digital-transformation/>.

GASPAR, W. Artificial Intelligence in Brazil still needs a strategy. CyberBRICS, 28 mar. 2022. Available at: <https://cyberbrics.info/artificial-intelligence-in-brazil-still-needs-a-strategy/>.

GENDRON, W. ChatGPT needs to 'drink' a water bottle's worth of fresh water for every 20 to 50 questions you ask, researchers say. Business Insider, 14 abr. 2023. Available at: <https://www.businessinsider.com/chatgpt-generative-ai-water-use-environmental-impact-study-2023-4>.

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

JARVENPAA, S. L.; IVES, B. Introducing transformational information technologies: the case of the World Wide Web technology. International Journal of Electronic Commerce, v. 1, n. 1, p. 95-126, 1996. Available at: <https://www.jstor.org/stable/27750802>.

LI, P. et al. Making AI Less "Thirsty": Uncovering and Addressing the Secret Water Footprint of AI Models. 2023. Available at: <https://arxiv.org/pdf/2304.03271.pdf>.

LIN, J. China invested US$290.8 billion in semiconductor projects between 2021-2022. DIGITIMES Asia, 27 jun. 2023. Available at: <https://www.digitimes.com/news/a20230627VL205/china-ic-manufacturing-semiconductor-chips+components.html>.

LUCCIONI, S. The mounting human and environmental costs of generative AI. Ars Technica, 12 abr. 2023. Available at: <https://arstechnica.com/gadgets/2023/04/generative-ai-is-cool-but-lets-not-forget-its-human-and-environmental-costs/>.

MINISTÉRIO DA CIÊNCIA, TECNOLOGIA E INOVAÇÃO (MCTI). MCTI anuncia revisão da Estratégia Brasileira de Inteligência Artificial. 11 dez. 2023. Available at: <https://www.gov.br/mcti/pt-br/acompanhe-o-mcti/noticias/2023/12/mcti-anuncia-revisao-da-estrategia-brasileira-de-inteligencia-artificial>.

MINISTÉRIO DE MINAS E ENERGIA. Brasil registra maior produção de energia limpa dos últimos 12 anos. 31 mar. 2023. Available at: <https://www.gov.br/mme/pt-br/assuntos/noticias/brasil-registra-maior-producao-de-energia-limpa-dos-ultimos-12-anos>.

MINISTÉRIO DO DESENVOLVIMENTO, INDÚSTRIA, COMÉRCIO E SERVIÇOS. Brasil ganha nova política industrial com metas e ações para o desenvolvimento até 2033. 22 jan. 2024. Available at: <https://www.gov.br/mdic/pt-br/assuntos/noticias/2024/janeiro/brasil-ganha-nova-politica-industrial-com-metas-e-acoes-para-o-desenvolvimento-ate-2033>.

NIC.br (Núcleo de Informação e Coordenação do Ponto BR). Conectividade Significativa: Propostas para medição e o retrato da população no Brasil. (2024). https://cetic.br/pt/publicacao/conectividade-significativa-propostas-para-medicao-e-o-retrato-da-populacao-no-brasil/

Pacotes “education” do Google e da Microsoft não contemplam lei brasileira de proteção de dados. Available at: <https://aberta.org.br/pacotes-education-nao-contemplam-lgpd/>.

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

SYNERGY RESEARCH GROUP. Cloud Market Gets its Mojo Back; AI Helps Push Q4 Increase in Cloud Spending to New Highs. fevereiro 2024. Available at: <https://www.srgresearch.com/articles/cloud-market-gets-its-mojo-back-q4-increase-in-cloud-spending-reaches-new-highs>.

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

TIC domicílios. Available at: <https://cetic.br/pt/pesquisa/domicilios/publicacoes/>.

Vipra, J. & Myers West, S. Computational Power and AI. AI Now Institute. (2023) https://ainowinstitute.org/publication/policy/compute-and-ai

Widder, David Gray and West, Sarah and Whittaker, Meredith, Open (For Business): Big Tech, Concentrated Power, and the Political Economy of Open AI. (2023). http://dx.doi.org/10.2139/ssrn.4543807

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.

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.

Introduction

African countries are taking steps in carving out their position as competitors in the development of Artificial Intelligence (AI). Policy positions are echoing the mantra 'AI made in Africa for Africa' (Smart Africa, 2021; AUDA NEPAD, 2024; African Union, 2024) as a way to signal the importance of African AI sovereignty characterised by the use of African data by Africans to develop products and services that can adequately address African challenges and meet African needs and demands, while also developing the African AI economy.

Earlier in 2024, the African Union Development Agency - New Partnership for Africa's Development (hereinafter "AUDA-NEPAD") took a decisive step in developing a continental wide white paper on AI[1] setting out African priorities and challenges in the wake of AI as well as proposing strategic pillars which can help Africa's AI ecosystem to thrive. Mid 2024, the African Union (AU) also adopted the first Continental AI Strategy[2]. The AU AI Strategy stresses the urgent need to develop the African AI economy. It highlights the importance of building capabilities to use AI for development. This includes building datasets, developing computing platforms for AI development, and nurturing AI skills and talent. The strategy also calls for local research, innovation and increased AI investments. It emphasises the need for AI governance and regulation, maximising AI benefits for socio-

[1] AUDA-NEPAD White Paper: Regulation and Responsible Adoption of AI in Africa. Towards achievement of Agenda 2063. AUDA NEPAD, February 2024. https://www.nepad.org/blog/taking-continental-leap-towards-technologically-empowered-africa-auda-nepad-ai-dialogue.

[2] Continental Artificial Intelligence Strategy: Harnessing AI for Africa's Development and Prosperity. AU, July 2024. https://au.int/en/documents/20240809/continental-artificial-intelligence-strategy.

economic development and cultural renaissance, minimising risk for responsible and secure AI in Africa, and regional and international cooperation on AI.

Prior to the AU adopting its policy position on AI, some African countries, through the Smart Africa Alliance[3] had initiated conversations on AI development in Africa. South Africa took over the country flagship on AI and developed a Blueprint on AI for Africa[4]. This blueprint highlights the importance of African AI sovereignty and calls for regional coordination of AI strategies, economic initiatives and policies.

South Africa has taken steps to position itself as a leader in AI development. The government has shown strong political commitment to advancing AI, with President Cyril Ramaphosa demonstrating this commitment through his direct involvement. In a key move, President Ramaphosa appointed a team of AI experts to form the Presidential Commission on the 4th Industrial Revolution (4IR). The 4IR Presidential Commission provided guidance to the country in the development of a strategy on 4IR, including AI (Presidential Commission on 4IR, 2019). The Presidential Commission on 4IR produced a report (Presidential Commission on 4IR, 2020) 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 or adoption of new policy and legislation; and establish 4IR strategy and implementation coordination council in the Presidency (*ibid.*). Some of the recommendations such as the establishment of the AI institute and the development of national policies were achieved within record time[5]. However, others require more time and significant investments.

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 (Hlomani, 2023).

[3] 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/.

[4] Blueprint: Artificial Intelligence for Africa. Smart Africa, 2021. https://smartafrica.org/knowledge/artificial-intelligence-for-africa/.

[5] For example, the Department of Communications and Digital Technologies (DCDT), the University of Johannesburg, and the Tshwane University of Technology created the AI Institute of South Africa in 2022. The DCDT also published the first draft of the National Data and Cloud Policy in 2021 for public comment and finally adopted a revised version in 2024.

South Africa recently adopted its National Data and Cloud Policy[6] (hereinafter "NDCP 2024") and National AI Policy Framework[7] which have the potential to extensively shape the future of AI development in South Africa. The AI Policy Framework is the first step in developing the country's National AI Policy. It identifies challenges that South Africa faces, various driving forces, aspirations and historical constraints shaping the development and implementation of AI in South Africa. For example, historical socio-economic inequalities, persistent digital divides and institutional resistance to change are some of the identified active barriers to AI development in the country. To take the country forward, the AI Policy Framework proposed strategic pillars similar to those outlined in the AU AI Strategy. The NDCP 2024 promotes access to data by SMMEs, startups, and citizens to drive innovation and develop digital solutions and tradable digital goods and services.

According to Belli, AI Sovereignty (hereafter "AIS") is not a universally defined concept (Belli, 2023). 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*" (*ibid.*). In this paper, we analyse the 8-point Key AI Sovereignty Enablers (KASE) (*ibid.*) 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.

3.1. Key AI Sovereignty Enablers (KASE)

3.1.1. Data governance

The different adages 'data is the new oil' or 'data is the new gold' are analogies (though not perfect analogies to capture the non-rivalrous and inexhaustible nature of data) 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.[8] The NDCP 2024 outlines the government's approach to data collection, storage, usage, and sharing, along with strategies for using cloud technologies to improve services and drive digital economic growth. There has been a major shift in policy interventions previously proposed in the Draft National Data and Cloud Policy of 2021 (hereinafter "NDCP 2021")[9] and the version which was adopted as the NDCP 2024. The NDCP 2021 focused on ownership of data generated within South Africa by the government and placed heavy emphasis on data localisation (see para 10.4.). After careful

[6] National Data and Cloud Policy. Department of Communications and Digital Technologies Notice 2533 of 2024.

[7] South Africa National Artificial Intelligence Policy Framework. Department of Communications and Digital Technologies. August 2024.

[8] 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.

[9] Draft National Policy on Data and Cloud. Department of Communications and Digital Technologies Notice 306 of 2021.

consideration of public input on the draft policy, the adopted NDCP 2024 is more progressive as it emphasises that given the digital infrastructural challenges, it may not be feasible for the government to own and manage data centres with its limited resource base. Intead, the NDCP 2024 supports and promotes data sharing as well as for government data to be stored in unified, cloud-enabled data centers, supported by redundancies in designated locations to meet business continuity requirements.

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.[10] 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[11] and those against big pharma (Information Regulator South Africa, 2023c). It remains to be seen how the Information Regulator (hereafter "IR") will handle some of the uses of AI, especially facial recognition software used in public spaces or use of AI in digital ID systems (Musoni et al, 2023). The IR 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 (Global Privacy Assembly, 2020a), the 2020 GPA Resolution on Facial Recognition Technology (Global Privacy Assembly, 2020b) or the 2022 GPA Resolution on Principles and Expectations for the Appropriate Use of Personal Information in Facial Recognition Technology (Global Privacy Assembly, 2020c). 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.

3.2. 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

[10] 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.

[11] The Information Regulator issued an enforcement notice against the South African Police Service (Information Regulator South Africa, 2023a); The Information Regulator also issued an enforcement notice against the Department of Justice and Constitutional Development where it is housed under (Information Regulator South Africa, 2023b).

Africans (Olorunju, 2022). There is a need to follow an African – Centred Approach[12] which includes incorporation of 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 (*ibid.*).

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 (Witwatersrand University, 2021). Government entities are also encouraged to procure services of software developers to develop proprietary software which will be owned by the government (Department of Public Service Administration, n.d.). 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 (Belli, 2023) of ubuntu.

3.3. 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 (Smart Africa, n.d.). Data centres play a significant role in providing the processing capacity, storage solutions and delivering an integrated AI infrastructure, applications and services. Under the NDCP, South Africa's policy position was to prepare 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 (South Africa, 2021). The approach it anticipated aimed 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 (Van der Berg, 2021; Razzano, 2021; Research ICT Africa, 2021). The public pushback may have contributed to South Africa changing its policy position under the NDCP 2021. For example, the NDCP 2021 had proposed the establishment of a High-Performance Computing and Data Processing Centre (HPCDPC) which will include processing and data

[12] 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.

facilities and cloud computing capacity and will consolidate existing public funded data centres, providing use-on-demand cloud services for State entities, national departments, provinces, municipalities, metros, universities, research centres, civil society organisations, and local businesses. However, this proposed intervention was left out in the adopted NDCP 2024. The NDCP 2024 promotes universal connectivity, supporting cloud-enabled data centers for government data and placing the State Information Technology Agency (SITA) in charge of overseeing data infrastructure sourcing, establishing service agreements, and enforcing security standards.

3.4. 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 (ICASA, 2023). 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 millions of those people using social media (Kemp, 2023). 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 (South Africa, 2020). 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 (Freedom House, 2021).

3.5. Reliable Electrical Power

Hyper scale data centres, 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 (South Africa, 2020; IMF, 2023). This has resulted in country wide daily power cuts or loadshedding undermining the country's economic recovery from COVID-19 (Lawlor, 2023). 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 centres and relying on power supply from independent power producers to reduce the total dependence on the strained national electricity grid (South Africa, 2020).

3.6. 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 (South Africa, 2020) 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 understand the fundamentals of AI, data and cloud computing, and how to access these to exploit economic opportunities (South Africa, 2021, Section 10.8.). 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 (Department of Basic Education, n.d.). The challenge for South Africa will be the availability of technical, financial and human resources to address the digital illiteracy rate.

3.7. 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.[13] The National Cybersecurity Policy Framework (South Africa, 2015) guides the implementation of cybersecurity initiatives and measures. The Cybercrimes Act (South Africa, 2020a) and the Critical Infrastructure Protection Act (South Africa, 2019a) 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 (De Stadler, Esselaar, 2015).

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 (*ibid.*).

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

3.8. 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

[13] 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, 2023).

deal with multiple regions and industries, including intellectual property and civil liability challenges, data protection, cybersecurity and ethical considerations (Smart Africa, n.d.). 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" (ACHPR, 2021).

South Africa is yet to develop a law specifically regulating AI. The existing legislation (like POPIA) can be applied to regulate the use of AI when processing personal data. The recently adopted data and AI policies focus almost exclusively on economic development and not on the appropriate use or ethical issues associated with AI (Ormond, n.d.). 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 (Musoni, 2023).

It is suggested that South Africa like many other African countries explore developing national AI Strategies to guide AI regulation adoption (Adams, 2022). Adams, citing Dutton, 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 (*ibid.*). The AI Policy Framework should be used as an opportunity for South Africa to flesh out the important issues that need to be covered when developing its national AI policy.

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 (UNESCO, 2022). The adopted AI policies should also advance gender equality through digital literacy and the inclusion of more women in digital spaces (Olorunju, 2022). South Africa can also take lessons from other African countries like Rwanda (Rwanda, 2022) and Egypt (Egypt, 2021), which have developed their own AI strategies and policies.

3.9. 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 (Ghosh, 2023). 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.

References

ADAMS, R. AI in Africa – Key Concerns and Policy Consideration for the Future of the Continent, 2022. Available at https://afripoli.org/ai-in-africa-key-concerns-and-policy-considerations-for-the-future-of-the-continent.

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. Available at: https://achpr.au.int/en/adopted-resolutions/473-resolution-need-undertake-study-human-and-peoples-rights-and-art.

AFRICAN CYBERCRIME OPERATIONS DESK. African cyberthreat assessment report: Cyberthreat trends outlook. Interpol, 2023.

AUDA-NEPAD. White Paper: Regulation and Responsible Adoption of AI in Africa. Towards achievement of Agenda 2063. February 2024. Available at https://www.nepad.org/blog/taking-continental-leap-towards-technologically-empowered-africa-auda-nepad-ai-dialogue.

AFRICAN UNION. Continental Artificial Intelligence Strategy: Harnessing AI for Africa's Development and Prosperity. AU, July 2024. Availabe at https://au.int/en/documents/20240809/continental-artificial-intelligence-strategy.

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.

Constitution of South Africa, Act 108 of 1996.

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

DE STADLER, E.; ESSELAAR, P. A Practical Guide to the Protection of Personal Information Act, 2015, p. 35.

DEPARTMENT OF BASIC EDUCATION. DBE to pilot draft curriculum on coding and robotics in schools. Available at: https://www.education.gov.za/CodingRoboticsPilot.aspx.

DEPARTMENT OF PUBLIC SERVICE ADMINISTRATION. Policy on Free and Open Source Software Use for South African Government. Available at: <https://www.gov.za/sites/default/files/gcis_document/201409/fosspolicy0.pdf>.

EGYPT. National AI Strategy. Available at: https://mcit.gov.eg/Upcont/Documents/Publications_672021000_Egypt-National-AI-Strategy-English.pdf.

FREEDOM HOUSE. South Africa: Freedom on the net, 2021. Available at: <https://freedomhouse.org/country/south-africa/freedom-net/2021>.

GHOSH, M. The case for AI foundation models as digital public infrastructure, Medium, 11 July, 2023. Available at: https://g-mainak.medium.com/the-case-for-ai-foundation-models-as-digital-public-infrastructure-3ea45896b5bf.

GLOBAL PRIVACY ASSEMBLY. Adopted Resolution On Accountability In The Development And Use Of Artificial Intelligence. 42nd Closed Session Of The Global Privacy Assembly, October 2020a. Available at: https://globalprivacyassembly.org/wp-content/uploads/2020/11/GPA-Resolution-on-Accountability-in-the-Development-and-Use-of-AI-EN.pdf.

GLOBAL PRIVACY ASSEMBLY. Adopted Resolution On Facial Recognition Technology. 42nd Closed Session Of The Global Privacy Assembly, October 2020b. Available at:

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

GLOBAL PRIVACY ASSEMBLY. Resolution on Principles and Expectations for the Appropriate Use of Personal Information in Facial Recognition Technology. 42nd Closed Session Of The Global Privacy Assembly, October 2020c. Available at: 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.

GRAVETT, W. H. Is the Dawn of the Robot Lawyer upon us? The Fourth Industrial Revolution and the Future of Lawyers. PER, v. 23, 2020. Available at: <http://www.scielo.org.za/scielo.php?script=sci_arttext&pid=S1727-37812020000100024&lng=en&nrm=iso>.

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.

HLOMANI, H. Why South Africa needs a more holistic and contextualized approach to AI regulation. Daily Maverick, 2023. Available at: https://www.dailymaverick.co.za/article/2023-05-23-why-south-africa-needs-a-more-holistic-and-contextual-approach-to-ai-regulation/.

ICASA. Icasa Publishes Three Final Radio Frequency Spectrum Assignment Plans For High-Demand Spectrum. 13 April, 2023. Available at: https://www.icasa.org.za/news/2023/icasa-publishes-three-final-radio-frequency-spectrum-assignment-plans-for-high-demand-spectrum.

IMF 'South Africa's Economy Loses Momentum Amid Record Power Cuts'. Available at: https://www.imf.org/en/News/Articles/2023/06/15/cf-south-africas-economy-loses-momentum-amid-record-power-cuts.

IMF. South Africa's Economy Loses Momentum Amid Record Power Cuts, 15 June, 2023. Available at: https://www.imf.org/en/News/Articles/2023/06/15/cf-south-africas-economy-loses-momentum-amid-record-power-cuts.

Information Regulator South Africa. Enforcement notice in terms of section 95 of the protection of personal information act 4 of 2013, 4 April, 2023a. Available at: https://inforegulator.org.za/wp-content/uploads/2020/07/ENFORCEMENT-NOTICE-SAPS-MATTER-04052363.pdf.

Information Regulator South Africa. Enforcement notice in terms of section 95 of the protection of personal information act 4 of 2013, 9 May, 2023b. Available at: https://inforegulator.org.za/wp-content/uploads/2020/07/ENFORCEMENT-NOTICE-SAPS-MATTER-04052363.pdf.

Information Regulator South Africa. Enforcement notice issued to Dis-Chem due to contravention of POPIA, 1 September, 2023c. Available at: https://inforegulator.org.za/wp-content/uploads/2020/07/FINAL-MEDIA-STATEMENT-ENFORCEMENT-NOTICE-ISSUED-TO-DISCHEM-PHARMACIES-LTD.pdf.

KEMP, S. Digital 2023: South Africa. Datareportal, 13 February, 2023. Available at: https://datareportal.com/reports/digital-2023-south-africa.

LAWLOR, P. SA's load shedding constraint and its impact on different economic sectors. Investec, March 02, 2023. Available at: https://www.investec.com/en_za/focus/economy/sa-s-load-shedding-how-the-sectors-are-being-affected.html.

MUSONI, M. Looking into the crystal ball: Artificial intelligence policy and regulation in Africa. Available at: <https://ecdpm.org/work/looking-crystal-ball-artificial-intelligence-policy-regulation-africa>.

MUSONI, M; DOMINGO, E; OGAH, E. Digital ID Systems in Africa: Challenges, risks and opportunities. ECDPM Discussion Paper 360. 1 December 2023. Available at https://ecdpm.org/work/digital-id-systems-africa-challenges-risks-and-opportunities

NICIS. Advancing High-Performance Computing in South Africa: The CHPC. Available at: https://www.nicis.ac.za/chpc/.

OLORUNJU, N. African Algorithmic Governance: Benefit of a Community-based Approach, 2022. Available at: <https://researchictafrica.net/2022/04/03/african-algorithmic-governance-benefit-of-a-community-based-approach/>.

ORMOND, E. Global To Local: South African Perspectives on AI Ethics Risks. Available at: https://ssrn.com/abstract=4240356.

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.

Presidential Commission on 4IR. 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.

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. Available at: <https://researchictafrica.net/wp/wp-content/uploads/2021/06/RIA_Submission_DATA_and_Cloud_Policy.pdf>.

RESOLUTION ON PRINCIPLES AND EXPECTATIONS FOR THE APPROPRIATE USE OF PERSONAL INFORMATION IN FACIAL RECOGNITION TECHNOLOGY. Available at: <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>.

RWANDA. National AI Policy. Available at: https://www.minict.gov.rw/index.php?eID=dumpFile&t=f&f=67550&token=6195a53203e197efa47592f40ff4aaf24579640e.

SA News. Government finalising national data, cloud policy, 13 April, 2023. Available at: https://www.sanews.gov.za/south-africa/government-finalising-national-data-cloud-policy.

SHUBHENDU, S.; VIJAY, J. Applicability of Artificial Intelligence in Different Fields of Life. **International Journal of Scientific Engineering and Research**, v. 1, n. 1, 2013.

SMART AFRICA ALLIANCE. Smart Africa member states each lead a specific project related to digital. Available at: <https://smartafrica.org/who-we-are/>.

SMART AFRICA. Blueprint: Artificial Intelligence for Africa. Available at: https://smartafrica.org/wp-content/uploads/2023/11/70029-eng_ai-for-africa-blueprint-min.pdf.

SOUTH AFRICA. Constitution of South Africa, Act 108 of 1996.

SOUTH AFRICA. Critical Infrastructure Protection Act 8 of 2019, 2019a.

SOUTH AFRICA. Cybercrimes Act 19 of 2020, 2020a.

SOUTH AFRICA. Industrial Policy Action Plan 2017/18 - 2019/20.

SOUTH AFRICA. National Cybersecurity Policy Framework. Government Gazette No. 39475 GN 609, 4 dez. 2015. Available at: https://www.gov.za/sites/default/files/gcis_document/201512/39475gon609.pdf.

SOUTH AFRICA. National Data and Cloud Policy. Government Gazette No. 44389 Government Notice 306, 1 April. 2021. Available at: https://www.gov.za/sites/default/files/gcis_document/202104/44389gon206.pdf.

SOUTH AFRICA. National Cloud and Data Policy. Government Gazette No. 50741 Government Notice 2544, 31 May 2024. Available at https://www.gov.za/sites/default/files/gcis_document/202406/50741gen2533.pdf

SOUTH AFRICA. National Digital and Future Skills Strategy. Government Gazette No. 43730, 30 ago. 2020b. Available at: https://www.gov.za/sites/default/files/gcis_document/202009/43730gen513.pdf.

SOUTH AFRICA. Presidential Commission on 4IR. Government No. Gazette 42388, 9 abr. 2019b. Available at: <https://www.gov.za/documents/presidential-commission-fourth-industrial-revolution-members-and-terms-reference-9-apr>.

SOUTH AFRICA. Presidential Commission on Fourth Industrial Revolution Report. Government Gazette No. 43834, 23 out. 2020c. Available at: https://www.gov.za/sites/default/files/gcis_document/202010/43834gen591.pdf.

SOUTH AFRICA. Report of the Presidential Commission on the 4th Industrial Revolution, 23 October, 2020d. Available at: https://www.gov.za/documents/notices/report-presidential-commission-4th-industrial-revolution-23-oct-2020.

TURING, A. Computing Machinery and Intelligence. Mind, v. 59, n. 236, 1950, p. 4337.

UNESCO. Landscape study of AI policies and use in Southern Africa: research report, 2022.

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

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.

WITWATERSRAND UNIVERSITY. AI-powered Algorithm released to detect the third wave in South Africa, 11 April, 2021. Available at: 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.

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 (AI) 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, such as the National Strategy on AI (Niti Aayog, 2018), as well as expert committees established by the Ministry of Electronics and Information Technology (Ministry of Electronics and Information Technology, Government of India, 2020b), and state-level interventions, including AI

industrial policies in Telangana (Government of Telangana, 2021) and Karnataka, and AI ethics and procurement frameworks by the Government of Tamil Nadu (Information Technology Department, Government of Tamil Nadu, 2020).[1]

In many of these policy documents, like the National Strategy on Responsible AI (National Strategy on Artificial Intelligence, n.d.), or the Department of Telecom's report on building an Indian AI Stack (Indian Artificial Intelligence Stack, 2021), **considerations of infrastructure** are paramount – they 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 (National Informatics Centre, n.d.), or through the incentives provided[2] to cloud service platforms to develop data centres in India, which have also been incorporated into policies like the Draft National Data Centre Policy (Ministry of Electronics and Information Technology, 2020c) and (draft) E-Commerce Policy (Ministry of Commerce, n.d.).

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,[3] 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 (Telecom Regulatory Authority of India, 2017).

However, infrastructural concerns are still prevalent across these areas, from the stability of electricity to the computational infrastructure required for training AI models (Husanjot Chahal et. al., 2021). There are clear disparities in the ability to access infrastructure required to develop or use AI applications, and increasing concentration of internet, 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 (Parsheera, Trehan, 2022). 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 (TRAI, 2023).

---

[1] 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

[2] 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* e.g., Government of Karnataka, 2022.

[3] 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, 2023.

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.[4] 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 (Ministry of Electronics and Information Technology, 2020a), which aim to make certain categories of 'non-personal data' available to public authorities as well as AI developers. Similarly, the Government of India's infrastructural interventions like those made in the National Health Stack (Niti Aayog, 2018) and financial Account Aggregator systems (India Stack, n.d.), intend to open up various forms of data for reusage in AI development supply chains, including, for example, through the facilitation of data exchange and consente mechanisms.

These proposals present some options on how datasets for AI development can be created outside the stronghold of big tech 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 (Veale, Binns, 2017), 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, 2023 (MEITY, 2023, Section 3).

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 (Telecom Regulatory Authority of India, 2023), 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 draft policy on Non-Personal Data Governance (which is concerned with regulating the availability and use of 'non-personal' 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

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

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 (Justice KS Puttaswamy v Union of India, 2017), 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 (Government of India, 2023).

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, India, 2021), or government procurement of algorithmic systems (as in the Tamil Nadu AI Policy, Information Technology Department, Government of Tamil Nadu, 2020).

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’.

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 (Fisher, Streinz, 2021), 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.

A 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 (Caplan et al., 2019; NIST, 2023).

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 (Dencik, 2019), outsourcing the work of labelling and moderation of the data used in AI

production to low-wage labour (Sarayu, 2021; Perrigo, 2023), used to hold sway over the livelihoods of populations affected by automation and shifts in technology-mediated skills (Cole, 2023), and making consequential decisions about people through the rarefied lenses of data science, big data and AI (Hildebrandt, 2019).

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, flows and relations 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 (Ricaurte, 2021).

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 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 (Walter, 2021), 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 (Lewis, 2020; CIFAR, n.d.).

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.

References

Natarajan, Sarayu; Mohamed, Suha; Mishra, Kushang; Taylor, Alex. Just and Equitable Data Labelling towards a Responsible AI Supply Chain. Aapti Institute, 2021. Available at: <https://aapti.in/blog/just-and-equitable-data-labelling/>.

PERRIGO, Billy. EXCLUSIVE: The $2 Per Hour Workers Who Made ChatGPT Safer. Time, 18 jan. 2023. Available at: <https://time.com/6247678/openai-chatgpt-kenya-workers/>.

Caplan, Robyn; Donovan, Joan; Hanson, Lauren; Matthews, Jeanna. ALGORITHMIC ACCOUNTABILITY: A PRIMER. Data and Society, 2019. Available at: <https://datasociety.net/library/algorithmic-accountability-a-primer/>.

NIST. AI Risk Management Framework 1.0. NIST, 2023. Available at: <https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf>.

CHAHAL, H. et al. Mapping India's AI Potential. Centre for Security and Emerging Technology, 2021.

COLE, M. (Infra)Structural Discontinuity: Capital, Labour, and Technological Change. Antipode, v. 55, p. 348, 2023.

DENCIK, L. et al. Exploring Data Justice: Conceptions, Applications and Directions. Information, Communication & Society, v. 22, p. 873, 2019.

DRAFT E-COMMERCE POLICY, 2019.

Non-Personal Data Policy, 2020.

FISHER, A.; STREINZ, T. Confronting Data Inequality. Columbia Journal of Transnational Law, v. 60, p. 829, 2021.

GOVERNMENT OF KARNATAKA. IndiAI. Available at: <https://indiaai.gov.in/ministries/government-of-karnataka?initiative=centre-of-excellence-in-data-sciences-and-artificial-intelligence>.

GOVERNMENT OF KARNATAKA. Karnataka Data Centre Policy, 2022. Available at: <https://itbtst.karnataka.gov.in/storage/pdf-files/Data%20Center%20Policy.pdf>.

GOVERNMENT OF TELANGANA. Telangana AI Framework, April 2021. Available at: <https://startup.telangana.gov.in/wp-content/uploads/2021/04/AI-framework.pdf>.

Lewis, Jason Edward, ed. 'Indigenous Protocol and Artificial Intelligence Position Paper', (2020).

CIFAR. Honolulu, Hawai'i: The Initiative for Indigenous Futures and the Canadian Institute for Advanced Research (CIFAR).

HILDEBRANDT, M. Privacy as Protection of the Incomputable Self: From Agnostic to Agonistic Machine Learning. Theoretical Inquiries in Law, v. 20, p. 83, 2019.

INDIA STACK. Available at: <https://indiastack.org/data.html>.

Government of INDIA. The Digital Personal Data Protection Act, 2023. (Act 22 of 2023).

Government of INDIA. The Information Technology (Intermediary Guidelines and Digital Media Ethics Code) Rules, 2021.

INDIAN ARTIFICIAL INTELLIGENCE STACK. Department of Telecommunications, 2021. Available at: <https://www.tec.gov.in/pdf/Whatsnew/ARTIFICIAL%20INTELLIGENCE%20-%20INDIAN%20STACK.pdf>.

INFORMATION TECHNOLOGY DEPARTMENT, GOVERNMENT OF TAMIL NADU. Tamil Nadu Safe and Ethical AI Policy, 2020. Available at: <https://elcot.in/sites/default/files/AIPolicy2020.pdf>.

JUSTICE K.S. Puttaswamy v Union of India, (2017) 10 SCC 1. Supreme Court of India.

MINISTRY OF COMMERCE, GOVERNMENT OF INDIA. Electronic Commerce in India: Draft National Policy Framework. Available at: <https://www.medianama.com/wp-content/uploads/Draft-National-E-commerce-Policy.pdf>.

MINISTRY OF ELECTRONICS AND INFORMATION TECHNOLOGY, GOVERNMENT OF INDIA. Report by the Committee of Experts on Non-Personal Data Governance Framework, 2020a. Available at: <https://static.mygov.in/rest/s3fs-public/mygov_160922880751553221.pdf>.

MINISTRY OF ELECTRONICS AND INFORMATION TECHNOLOGY, GOVERNMENT OF INDIA. Artificial Intelligence Committee Reports, 2020b. Available at: <https://www.meity.gov.in/artificial-intelligence-committees-reports>.

MINISTRY OF ELECTRONICS AND INFORMATION TECHNOLOGY, GOVERNMENT OF INDIA. Draft Data Centre Policy, 2020c. Available at: <https://www.meity.gov.in/writereaddata/files/Draft%20Data%20Centre%20Policy%20-%2003112020_v5.5.pdf>.

NATIONAL INFORMATICS CENTRE. National Cloud. Available at: <https://www.nic.in/servicecontents/national-cloud/>.

NITI Aayog. Government of India. NATIONAL STRATEGY ON ARTIFICIAL INTELLIGENCE.. Available at: <https://niti.gov.in/national-strategy-artificial-intelligence>.

NITI AAYOG. Government of India. National Health Stack – Strategy and Approach, 2018. Available at: <https://abdm.gov.in:8081/uploads/NHS_Strategy_and_Approach_1_89e2dd8f87.pdf>.

PARSHEERA, S.; TREHAN, V. A Structural Analysis of the Mobile Telecommunications Market: Exploring the Jio Effect, 2022. Available at: <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/>.

RICAURTE, P. Data epistemologies, the coloniality of power, and resistance. Television & New Media, v. 20, n. 4, p. 350, 2019. FISHER, A.; STREINZ, T. Confronting data inequality. Colum. J. Transnat'l L., v. 60, p. 829, 2021.

MEITY. Digital Personal Data Protection Act, 2023. Available at: <https://www.meity.gov.in/writereaddata/files/Digital%20Personal%20Data%20Protection%20Act%202023.pdf>.

TELECOM REGULATORY AUTHORITY OF INDIA. Recommendations on Leveraging Artificial Intelligence and Big Data in Telecommunication Sector, jul. 2023. Available at: <https://www.trai.gov.in/sites/default/files/Recommendation_20072023_0.pdf>.

TELECOM REGULATORY AUTHORITY OF INDIA. Recommendations on Network Neutrality, 28 nov. 2017. Available at: <https://www.trai.gov.in/sites/default/files/Recommendations_NN_2017_11_28.pdf>.

TELECOM REGULATORY AUTHORITY OF INDIA. Telecom Subscription Data, Q1 2023. Available at: <https://www.trai.gov.in/sites/default/files/PR_No.46of2023_0.pdf>.

VEALE, M.; BINNS, R. Fairer Machine Learning in the Real World: Mitigating Discrimination without Collecting Sensitive Data. 2017. Available at: <https://journals-sagepub-com.libproxy.ucl.ac.uk/doi/full/10.1177/2053951717743530>.

WALTER, M. et al. Indigenous Data Sovereignty in the Era of Big Data and Open Data. Australian Journal of Social Issues, v. 56, p. 143-156, 2021.

PART 2:
WHAT DO AI TRANSPARENCY AND AI ACCOUNTABILITY MEAN?

5. Broadening the Horizon: New Concepts for AI Regulation

Rolf H. 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 and is reflected in many international instruments (such as the OECD AI Guidelines of May 2019). 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" (BRANDEIS, 1914, p. 92).

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 (WEBER, 2023; UNESCO, 2023).[1] This essay questions the assumption that transparency is sufficient to combat AI challenges and proposes to apply additional regulatory models.

[1] This article is partly based on Weber (2023); this publication analyzing a comparable digital appearance will not be cited anymore further on. Very recently, transparency has again been taken up as a key objective in UNESCO (2023).

Often transparency is differentiated into three main pillars, namely (i) procedural transparency, (ii) decision-making transparency and (iii) substantive transparency:[2]

> *(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.

5.1.2. 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 analyses. Echoing these voices, transparency is partly seen as policy panacea.[3] AI data processing is exposed to false binaries between secrecy and openness, to strategic occlusions and to market-dominant behaviour of big enterprises; these factors influencing the algorithmic matching results can lead to power imbalances.

5.1.3. 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.[4] 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.

---

2 See Rolf H. Weber, Shaping Internet Governance: Regulatory Challenges, Zurich 2009, 121.

3 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.

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

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".[5] If the information is clear and straightforward, the addressee will be able to fully understand it (so-called comprehensibility).[6] 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).

5.1.4. 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".[7] Apart from the hidden costs caused by such kind of disclosure (for example detailed information obligations in the Artificial Intelligence Act of the EU[8]), Ben-Shahar & Schneider argue that the mandated disclosure would exacerbate inequality, impair consumers' decisions and deter lawmakers from adopting better regulations;[9] 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".[10]

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.

5.1.5. Information Overload

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

6 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.

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

8 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 late fall 2023.

9 See Ben-Shahar/Schneider (supra note 147); for a more detailed discussion see Weber (supra note 146), 75 and 77–78.

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

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:[11]

- 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:[12] (i) Over-information consumes working and leisure time on both sides of an informational relationship. (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;[13] as a consequence, the regulator should recognize that "less is more" even if it cannot be excluded that "less is not enough".[14]

5.2. Accountability

5.2.1. 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.[15] 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.

---

11 See Weber (supra note 146), 79–80 with further references.

12 See Niklas Luhmann, Die Gesellschaft der Gesellschaft, Frankfurt 1997, 1090, 1097 and 1102.

13 Weber (supra note 146), 79–80.

14 Ben-Shahar/Schneider (supra note 150), 647.

15 Weber (supra note 141), 133 with further references; accountability also is an important objective in the recent UNESCO Guidelines (supra note 141), 10, 21, 27/28, 48/49.

As a fundamental principle, accountability concerns itself with power and power implies responsibility. Therefore, accountability can be framed among three elements,[16] 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 of the decision-making processes should be acceptable to civil society in general.[17]

5.2.2. 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.[18] However, accountability should extend beyond oversight of algorithms and behavioural conduct. In addition, behaviour cannot be reduced to conduct since tight integration of data collections and targeted "intervention" in form of "surveillance capitalism"[19] has produced a market form that is unimaginable outside the digital milieu. As a consequence, the rising power 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.[20]

A specific challenge to accountability is the occurrence of opacity[21] 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".[22] 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.[23] 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 centres and source code, the above discussed comprehensibility would not be straightforward in view of complex code

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

17 Rieder/Hofmann (supra note 143), 6.

18 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.

19 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.

20 To the autonomy and power elements in the accountability context see also Weber (supra note 156), 71.

21 Rieder/Hofmann (supra note 143), 6–7.

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

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

designs and involved machine learning. In addition, the existence of different programming languages and execution environments adds further complications.[24]

5.3. Alternative Approaches and Regulatory Models

5.3.1. 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 are 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.[25] 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.[26] 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 analysed. Insofar soft law instruments developed by the concerned stakeholders of civil society merit special attention.

5.3.2. Auditability

An improvement of transparency and accountability can be achieved by extended auditability requirements if the respective provisions overcome an insufficient understanding of algorithms and platform architectures.[27] 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:[28] (i) The creation of an intermediary (public or private sector entity) that audits

---

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

25 Rieder/Hofmann (supra note 143), 5.

26 Weber (supra note 144), 70.

27 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.

28 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.

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.[29] Such an institution could be created within the context of the recently adopted EU Digital Services Act (DSA).[30] 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.[31]

5.3.3. 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.[32] 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".[33] 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.[34] The concept of observability seeks to develop concrete actions in respect of (i) how people need to be treated in the digitized environment,

29 Wagner/Kuklis (supra note 168), 174.

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

31 Wagner/Kuklis (supra note 168), 174–175.

32 A broad discussion of the observability concept is offered by Rieder/Hofmann (supra note 143), 9–18.

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

34 Rieder/Hofmann (supra note 143), 9–10.

(ii) how connections between participants are made and structured, and (iii) which outcomes should be achievable.[35]

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.[36] 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 freedom of expression and the freedom of information, fostering cultural and political diversity throughout the whole society.[37]

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.[38] 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.[39]

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.[40]

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.[41] AI providers should become subject to public interest requirements as a normative benchmark; elements could consist of risk control measures, auditability reviews, behavioural rules and strict responsibility obligations.

5.4. Soft Law as Alternative Formal Avenue

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

36 Rieder/Hofmann (supra note 143), 9–10.

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

38 Rieder/Hofmann (supra note 143), 13–14.

39 Von Djick (supra note 177), 3.

40 Bovens (supra note 173), 450.

41 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.

As far as the normative rule-making models are concerned, a fresh thinking appears to be necessary. Existing and future governmental 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,[42] 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.[43] The respective structure could look as follows:

Laws (by States)
Guidelines (by international borders
Co-regulation
Self-regulation
Voluntary compliance

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;[44] the "Model Framework" refers, as the name says, to digital platforms, but AI services have parallel characteristics in many 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.[45]

The "Model Framework" proposes standardized and shared rules:[46] 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 in their algorithms. As mentioned, these principles could comparably be designed in respect of the AI services providers' commitments.

---

42 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.

43 Weber (supra note 182), 27–28 with further references.

44 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.

45 Belli et al. (supra note 184), 7.

46 Belli et al. (supra note 184), 7.

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. In addition, information about the complaints proceedings (appeals received, treated, accepted and rejected) must be made available.

5.5. 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.[47]

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 and for improving interoperability in AI risk management.[48]

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:[49] (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 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.

References

ANNANY, M.; CRAWFORD, K. Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. New Media & Society, v. 20, 2018, p. 973–974.

47 See also Rieder/Hofmann (supra note 143), 23.

48 For further details see Weber (supra note 142), 140–143, 147. To the interoperability issue see OECD, Common Guideposts to Promote Interoperability in A Risk Management, Paris, November 2023, https://www.oecd.org/publications/common-guideposts-to-promote-interoperability-in-ai-risk-management-ba602d18-en.htm.

49 See Belli et al. (supra note 184), 6–7, in respect of digital platforms.

BELLI, L.; CURZI, Y.; ALMEIDA, C.; COUTO, N.; RADU, R.; WEBER, R. H.; BROWN, I. Towards Meaningful and Interoperable Transparency for Digital Platforms. UN IGF, 2022. Available at: <https://www.intgovforum.org/en/filedepot_download/57/23886>.

BEN-SHAHAR, O.; SCHNEIDER, C. E. More Than You Wanted to Know: The Failure of Mandated Disclosure. Princeton, 2014.

BEN-SHAHAR, O.; SCHNEIDER, C. E. The Failure of Mandated Disclosure. University of Pennsylvania Law Review, v. 159, 2011, p. 647, 748.

BOVENS, M. Analysing and Assessing Accountability: A Conceptual Framework. European Law Journal, v. 13, 2007, p. 447, 450.

BRANDEIS, L. The Other People's Money and How the Bankers Use It. New York, 1914.

BURRELL, J. How the machine 'thinks': Understanding opacity in machine learning algorithms. Big Data & Society, v. 3, 2016, p. 1–2.

DIJCK, J. van. Governing digital societies: Private platforms, public values. Computer Law & Security Review, v. 36, 2020, p. 1, 3.

DIJCK, J. van; POELL, T.; WAAL, M. de. The Values in a Connective World. Oxford, 2018, p. 158.

DOURISH, P. Algorithms and their others: Algorithmic cultures in context. Big Data & Society, v. 3, 2016, p. 1, 4.

EUROPEAN COMMISSION. Proposal for a Regulation laying down harmonised rules on artificial intelligence (Artificial Intelligence Act, AIA) of April 21, 2021. COM (2021) 206 final. Available at: <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206>.

EUROPEAN UNION. General Data Protection Regulation (GDPR) 2016/679 of April 2016. OJ 2016 L 119 of 4 May 2016.

EUROPEAN UNION. Regulation 2022/2065 of 19 October 2022 on a Single Market for Digital Services. OJ 277 L 1 of 27 October 2022.

KOENE, A. et al. A governance framework for algorithmic accountability and transparency. European Parliamentary Research Service Study, abr. 2019, p. 31–32.

LUHMANN, N. Die Gesellschaft der Gesellschaft. Frankfurt, 1997, p. 1090, 1097, 1102.

NAPOLI, P. M. Social media and the public interest: Governance of new platforms in the realm of individual and algorithmic gatekeepers. Telecommunications Policy, v. 39, 2015, p. 751 et seq.

OECD. Common Guideposts to Promote Interoperability in A Risk Management. Paris, nov. 2023. Available at: <https://www.oecd.org/publications/common-guideposts-to-promote-interoperability-in-ai-risk-management-ba602d18-en.htm>.

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

RIEDER, B.; HOFMANN, J. Towards platform observability. Internet Policy Review, v. 9, 2020, p. 1, 3–6. Available at: <https://doi.org/10.14763/2020.4.1535>.

UNESCO. Guidelines for the Governance of Digital Platforms. Paris, set. 2023, p. 19, 21, 25, 42-45. Available at: <https://www.unesco.org/en/articles/guidelines-governance-digital-platforms>.

WAGNER, B.; KUKLIS, L. Establishing Auditing Intermediaries to Verify Platform Data. In: MOORE, M.; TAMBINI, D. (Eds.). Regulating Big Tech. Oxford, 2021, p. 169, 172–173.

Weber, R. H. Transparency on Digital Platforms, Weblaw Jusletter, August 31, 2023

WEBER, R. H. From Disclosure to Transparency in Consumer Law. In: MATHIS, K.; TOR, A. (Eds.). Consumer Law and Economics. Cham, 2021, p. 73, 79–81.

WEBER, R. H. Internet Governance at the Point of No Return. Zurich, 2021, p. 70.

WEBER, R. H. Sectoral Self-Regulation as a Viable Tool. In: MATHIS, K.; TOR, A. (Eds.). Law and Economics of Regulation. Cham, 2021, p. 5, 26–27.

WEBER, R. H. Shaping Internet Governance: Regulatory Challenges. Zurich, 2009, p. 121.

WEBER, R. H. The Disclosure Dream – Towards a New Transparency Concept in Consumer Law. EuCML, 2023, p. 67–68.

ZUBOFF, S. The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power. New York, 2019, p. 15.

Pre-print version of Belli L. Gaspar. W.B. The Quest for AI Sovereignty, Transparency and Accountability. Springer-Nature. (2025).

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 and assurance, incentivisation, efficacy and accountability. To support this framework, regulation will need to require the use of various transparency and assurance 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.

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,[1] and more broadly how regulation can be applied to information society services with complex supply chains suffering from a "many hands" problem.[2] Based on a rapid review of academic and grey literature, this article analyses which actors should be primarily responsible for different risks in 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 and assurance, 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 6.2 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 6.3 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.

6.1. 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,

---

[1] 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/

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

we provide an initial conceptual framework that regulators can follow to determine where responsibilities might apply, which relies on four principles:

- **Transparency and assurance:** 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.

6.1.1. Transparency and assurance

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 and providing assurance elements in the chain meet claimed standards.

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'.[3]
- **Certifications, audits, impact assessments, technical standards and similar mechanisms**, which give organisations reliable evidence on and assurance of the trustworthiness of AI systems.[4] 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

[3] ibid.

[4] 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.

also develop voluntary sectoral codes of conduct, building on those envisaged in the GDPR's Articles 40 and 41, and developing standards for certifications.[5]

- **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.'[6]
- **Mechanisms for reporting and remedying faults.** Researchers from Stanford's Human-Centred 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 users' including deployers or end users whose products and services rely on that model.[7]

More broadly, it may be most efficient for a government body to play a cross-sectoral role for information-sharing and learning.[8] 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.'[9] 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

---

[5] 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.

[6] 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.

[7] 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>.

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

[9] 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.

and ethicists create standards that allow us to hold flawed or biased AI accountable under the law.'[10] However, some researchers have noted the limitations of current XAI approaches, which can be brittle and change over time.[11]

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.

6.1.2. 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.[12]

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'.[13]

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'.[14]

6.1.3. Efficacy up and down the AI supply chain

[10] 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>.

[11] 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;

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

[13] 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.

[14] ibid.

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.[15] 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 those users 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 also any use case specific quality management process, technical documentation and logging, as well as any additional robustness and accuracy testing.'[16] This is because downstream deployers are closer in proximity to the final context in which the system is operating.

6.1.4. 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.'[17]

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.[18] Japan's government is encouraging this by issuing interpretive guidance on AI contracts.[19] 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 eventually become the providers, as they will place the high-risk AI systems on the market.'[20]

---

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

[16] 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.

[17] ibid.

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

[19] 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>.

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

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.[21]

6.2. 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 '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.'[22] 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.'[23]

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.

3. A single failure by the developer 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.'[24]

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.

---

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

[22] Engler and Renda (n 1).

[23] 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.

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

6.2.1. 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,[25] 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."[26] And already, "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."[27]

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'.[28] This would mean focusing regulatory attention on the large providers of these foundational models.

6.2.2. 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'.[29]

---

[25] 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.

[26] 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.

[27] 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.

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

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

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'.[30]

Locating responsibility with GPAI developers higher up the supply chain would enable them to 'control several levers that might 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.'[31]

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.'[32]

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. With one 2023 US survey[33] finding that 37% of marketing and advertising firms, and 35% of technology firms, had already adopted the technology, FTC rules will have an economy-wide impact. And 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'.[34] 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) and the UK Online Safety Act.[35] 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

---

[30] N Kolt, 'Algorithmic Black Swans' (2023) 101 Washington University Law Review 33.

[31] 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.

[32] 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.

[33] Fishbowl (2023)

[34] Widder and Nafus (n 14).

[35] 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.

for platforms under the Online Safety Act.[36] 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.[37]

The European Parliament has proposed tailored requirements for GPAI[38], 'foundation models'[39] and 'generative AI'[40]. It conceptualises foundation models and generative AI as sub-categories of GPAI, and set different rules for each:

1. 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.
2. Foundation model providers will have 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.
3. 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.

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.

6.3. 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

---

[36] 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.

[37] 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.

[38] "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"

[39] "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"

[40] 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"

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'[41]).

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.[42] (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.'[43])

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 of the model in a certain way, but it remains unclear how viable this method is as a remedy for preventing misuse.[44] 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."'[45]

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.

6.3.1. The challenges of open-source

Open-source GPAI projects play two key roles:

---

[41] DG Widder, S West and M Whittaker, 'Open (For Business): Big Tech, Concentrated Power, and the Political Economy of Open AI' (17 August 2023) <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4543807>. Accessed 6 April 2024.

[42] 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.

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

[44] 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.

[45] Widder and Nafus (n 14).

- '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.'[46]

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 generation system.[47] However, it is likely such contributions will be in the interests of the companies concerned.[48] An expert review for the European Commission found that platforms generally shape innovation within their own ecosystems to bolster their business models,[49] 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.[50]

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[51] 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[52] and which is now made available under licences with significant restrictions and lacking important information.[53] 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".[54] 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

---

[46] 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/>.

[47] 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.

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

[49] 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.

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

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

[52] 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.

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

[54] 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.

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 core model logics to the large companies who have the resources to create them from scratch."[55]

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[56]).
- 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[57] 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.[58]

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.[59]

Recognising this, advocates for regulation of 'frontier' (most capable) AI systems have suggested using controls on the sale of specialised processors (from companies such as Nvidia) necessary to create these models to enforce 'safe and ethical uses of the technology'.[60] 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.

---

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

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

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

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

[59] 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.

[60] 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.

6.4. 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."[61]

Transparency and assurance mechanisms like model cards, datasheets, audits, 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 competitiveness and safety.[62]

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.

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.[63]

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

---

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

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

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

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.

References

ACCESS NOW et al. Call for Better Protections of People Affected at the Source of the AI Value Chain, 25 out. 2022. Available at: <https://futureoflife.org/wp-content/uploads/2022/10/Civil-society-letter-GPAIS-October-2022.pdf>. Acesso em: 21 mar. 2023.

ADA LOVELACE INSTITUTE. Approaches to government monitoring of the AI landscape. Internal briefing for DCMS, 2023.

AHMED, N.; WAHED, M.; THOMPSON, N. C. The Growing Influence of Industry in AI Research. Science, v. 379, 2023, p. 884.

ATLESON, M. Chatbots, Deepfakes, and Voice Clones: AI Deception for Sale. Federal Trade Commission Business Blog, 20 mar. 2023. Available at: <https://www.ftc.gov/business-guidance/blog/2023/03/chatbots-deepfakes-voice-clones-ai-deception-sale>. Acesso em: 22 mar. 2023.

CENTRE FOR DATA ETHICS AND INNOVATION. The Roadmap to an Effective AI Assurance Ecosystem - Extended Version. 2021. Available at: <https://www.gov.uk/government/publications/the-roadmap-to-an-effective-ai-assurance-ecosystem/the-roadmap-to-an-effective-ai-assurance-ecosystem-extended-version>. Acesso em: 11 mar. 2023.

COBBE, J.; SINGH, J. Artificial Intelligence as a Service: Legal Responsibilities, Liabilities, and Policy Challenges. Computer Law & Security Review, v. 42, 2021, p. 105573, 43.

COBBE, J.; VEALE, M.; SINGH, J. Moving beyond "Many Hands": Accountability in Algorithmic Supply Chains. Proceedings of Fairness, Accountability and Transparency '23, 2023, p. 9.

CONTRACTOR, D. et al. Behavioral Use Licensing for Responsible AI. 2022 ACM Conference on Fairness, Accountability, and Transparency, 2022. Available at: <https://dl.acm.org/doi/10.1145/3531146.3533143>. Acesso em: 24 mar. 2023.

DARLEY, J. M.; LATANE, B. Bystander Intervention in Emergencies: Diffusion of Responsibility. Journal of Personality and Social Psychology, v. 8, 1968, p. 377.

DE BRUIJN, H.; WARNIER, M.; JANSSEN, M. The perils and pitfalls of explainable AI: Strategies for explaining algorithmic decision-making. Government Information Quarterly, v. 39, n. 2, 2022, p. 101666. Available at: <https://doi.org/10.1016/j.giq.2021.101666>.

DIFFIE, W.; LANDAU, S. Privacy on the Line. Updated and Expanded Edition. Random House, 2010. Available at: <https://www.penguinrandomhouse.com/books/654750/privacy-on-the-line-updated-and-expanded-edition-by-whitfield-diffie-and-susan-landau/>. Acesso em: 12 mar. 2023.

DONELAN, M.; PARKINSON, L. Online Safety Bill 2023. European Parliament and Council of the European Union. Digital Services Act, art. 34.

EDWARDS, L.; VEALE, M. Slave to the Algorithm? Why a "Right to an Explanation" Is Probably Not the Remedy You Are Looking For. Duke Law & Technology Review, v. 16, p. 70–80, 2017.

ENGLER, A. Early Thoughts on Regulating Generative AI like ChatGPT. 16 fev. 2023. Available at: <https://www.brookings.edu/blog/techtank/2023/02/16/early-thoughts-on-regulating-generative-ai-like-chatgpt/>. Acesso em: 21 fev. 2023.

ENGLER, A. The EU's Attempt to Regulate Open-Source AI Is Counterproductive. Brookings Institution TechTank, 24 ago. 2022. Available at: <https://www.brookings.edu/blog/techtank/2022/08/24/the-eus-attempt-to-regulate-open-source-ai-is-counterproductive/>.

ENGLER, A.; RENDA, A. Reconciling the AI Value Chain with the EU's Artificial Intelligence Act. Centre for European Policy Studies, 2022, p. 2–3. Available at: <https://www.ceps.eu/ceps-publications/reconciling-the-ai-value-chain-with-the-eus-artificial-intelligence-act/>.

EZRACHI, A.; STUCKE, M. E. Digitalisation and Its Impact on Innovation. European Commission DG Research and Innovation, 2020, p. 978-92-76-17462-2. Available at: <https://research-and-innovation.ec.europa.eu/knowledge-publications-tools-and-data/publications/all-publications/digitalisation-and-its-impact-innovation_en>. Acesso em: 21 mar. 2023.

FISHBOWL. ChatGPT Sees Strong Early Adoption In The Workplace, 17 jan. 2023. Available at: < https://www.statista.com/statistics/1361251/generative-ai-adoption-rate-at-work-by-industry-us/>. Accesso em: 23 oct. 2023.

HACKER, P.; ENGEL, A.; LIST, T. Understanding and Regulating ChatGPT, and Other Large Generative AI Models: With input from ChatGPT. Verfassungsblog, 20 jan. 2023. Available at: <https://verfassungsblog.de/chatgpt/>. Acesso em: 20 jan. 2023.

HELBERGER, N.; DIAKOPOULOS, N. ChatGPT and the AI Act. Internet Policy Review, v. 12, 2023. Available at: <https://policyreview.info/essay/chatgpt-and-ai-act>. Acesso em: 22 fev. 2023.

HELBERGER, N.; PIERSON, J.; POELL, T. Governing Online Platforms: From Contested to Cooperative Responsibility. The Information Society, v. 34, p. 1, 2018.

KAK, A.; WEST, S. M. AI Now 2023 Landscape: Confronting Tech Power. AI Now Institute, 2023, p. 17. Available at: <https://ainowinstitute.org/2023-landscape>. Acesso em: 21 maio 2023.

KOLT, N. Algorithmic Black Swans. Washington University Law Review, v. 101, 2023, p. 33.

KORTZ, M.; DOSHI-VELEZ, F. Accountability of AI Under the Law: The Role of Explanation. Berkman Klein Center, 2017. Available at: <https://cyber.harvard.edu/publications/2017/11/AIExplanation>.

LI, Y. et al. GLIGEN: Open-Set Grounded Text-to-Image Generation. 17 abr. 2023. Available at: <http://arxiv.org/abs/2301.07093>. Acesso em: 6 mar. 2023.

LIANG, P. et al. The Time Is Now to Develop Community Norms for the Release of Foundation Models. Stanford University Human-Centered Artificial Intelligence, 17 maio 2022. Available at: <https://hai.stanford.edu/news/time-now-develop-community-norms-release-foundation-models>.

MATWYSHYN, A. Hacking Speech: Informational Speech and the First Amendment. Northwestern University Law Review, v. 107, p. 795, 2013.

MEIT EXPERT GROUP. Governance Guidelines for Implementation of AI Principles Ver. 1.1. Japan Ministry of Economy, Trade and Industry, 2021, p. 35. Available at: <https://www.meti.go.jp/english/press/2022/0128_003.html>.

OVERSIGHT OF A.I.: Rules for Artificial Intelligence. Available at: <https://www.judiciary.senate.gov/download/2023-05-16-testimony-altman>. Acesso em: 21 maio 2023.

PATEL, D.; AHMAD, A. Google "We Have No Moat, And Neither Does OpenAI". semianalysis, 4 maio 2023. Available at: <https://www.semianalysis.com/p/google-we-have-no-moat-and-neither>. Acesso em: 19 maio 2023.

RICHTER, F. Amazon, Microsoft & Google Dominate Cloud Market. Statista Infographics, 23 dez. 2022. Available at: <https://www.statista.com/chart/18819/worldwide-market-share-of-leading-cloud-infrastructure-service-providers>. Acesso em: 21 mar. 2023.

SCHOONEWILLE, M. et al. Introduction New Algorithm Regulator and Implications for Financial Sector. Lexology, 5 jan. 2023. Available at: <https://www.lexology.com/library/detail.aspx?g=3e71f01b-2cb7-4294-b8f2-68ea2ab67261>. Acesso em: 20 jan. 2023.

SOLAIMAN, I. The Gradient of Generative AI Release: Methods and Considerations. Proceedings of Fairness, Accountability and Transparency '23, 2023. Available at: <http://arxiv.org/abs/2302.04844>. Acesso em: 25 fev. 2023.

TALESH, S. Cyber Insurance and Cybersecurity Policy: An Interconnected History. Lawfare, 4 nov. 2022. Available at: <https://www.lawfareblog.com/cyber-insurance-and-cybersecurity-policy-interconnected-history>. Acesso em: 23 mar. 2023.

URBINA, F. et al. Dual Use of Artificial-Intelligence-Powered Drug Discovery. Nature Machine Intelligence, 2022, p. 189.

VINCENT, J. OpenAI Co-Founder on Company's Past Approach to Openly Sharing Research: "We Were Wrong". The Verge, 15 mar. 2023. Available at: <https://www.theverge.com/2023/3/15/23640180/openai-gpt-4-launch-closed-research-ilya-sutskever-interview>. Acesso em: 24 mar. 2023.

VINCENT, J. OpenAI Co-Founder on Company's Past Approach to Openly Sharing Research: "We Were Wrong". The Verge, 15 mar. 2023.

WATERS, R. US Should Use Chip Leadership to Enforce AI Standards, Says Mustafa Suleyman. Financial Times, 1 set. 2023. Available at: <https://www.ft.com/content/f828fef3-862c-4022-99d0-41efbc73db80>. Acesso em: 8 set. 2023.

WHITTAKER, M. The Steep Cost of Capture. Interactions, v. 28, p. 50, 2021.

WIDDER, D. G.; NAFUS, D. Dislocated Accountabilities in the AI Supply Chain: Modularity and Developers' Notions of Responsibility. Big Data & Society, 2023. Available at: <http://arxiv.org/abs/2209.09780>. Acesso em: 17 jan. 2023.

WIDDER, D. G.; WEST, S.; WHITTAKER, M. Open (For Business): Big Tech, Concentrated Power, and the Political Economy of Open AI, 2023, p. 18. Available at: <https://papers.ssrn.com/abstract=4543807>. Acesso em: 7 set. 2023.

WIDDER, D. G.; WEST, S.; WHITTAKER, M. Open (For Business): Big Tech, Concentrated Power, and the Political Economy of Open AI. 17 ago. 2023. Available at: <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4543807>. Acesso em: 6 abr. 2024

Pre-print version of Belli L. Gaspar. W.B. The Quest for AI Sovereignty, Transparency and Accountability. Springer-Nature. (2025).

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.

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.[1] In the past decade, governments worldwide have intensified their scrutiny of the AI landscape, driven by the impetus to establish regulatory frameworks. Notably, the EU AI Act has assumed a pioneering role in delineating comprehensive regulations that encompass even software.[2] 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.[3]

[1] 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/

[2] 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

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

GenAI has permeated diverse spheres of human existence be it business, education, or recreation.[4] 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.[5] This information could potentially be harnessed for training purposes by the platform, prompting apprehensions within the business community.[6] Concurrently, labour commissions express concerns about its ramifications on employment dynamics.[7] Educational institutions, in turn, harbour reservations regarding the preservation of academic integrity, given GenAI's influence on student submissions.[8]

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 landscape, particularly for differently-abled members of our society.[9] 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.[10] It is pivotal to recognise that many GenAI platforms are constructed upon publicly available, and at times intellectually protected, information[11]. 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

---

[4] 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.

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

[6] 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.

[7] 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/

[8] 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

[9] 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

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

[11] 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

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 the socio-technical systems theory[12]. 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.

7.1. Research Methodology

7.1.1. 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.

7.1.2. 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.[13] Given GenAI's unique attributes and its profound societal implications, a dedicated privacy compliance framework is imperative to ensure its responsible evolution.

---

[12] 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

[13] 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.

7.2. 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.[14]
- **Outcome:** Feedback will guide the refinement of the initial framework, ensuring its practicality and applicability.

Stage 3: Empirical Analysis

- **Scope:** A 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?

5. Methodological Considerations

- **Doctrinal Approach:** The first stage involves a doctrinal research approach, investigating global privacy law resources, inclusive of statutes, regulations, case

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

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.

7.3. Deploying the Foundational Privacy Principles for Gen AI Platforms

7.3.1. 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.[15] The scholars delineated the primary stages – design, develop, and deploy – and further segmented these into 19 nuanced sub-stages. Such a comprehensive 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

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

explainability, responsibility, accountability, data documentation practices, evaluation, and monitoring, is mirrored in aspects of the CDAC framework'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.[16] 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.

7.4. 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[17], 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[18] 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

[16] 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.

[17] 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

[18] 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

At the very core of the GenAI lifecycle – encompassing design, development, and deployment stages – lies the principle of proactive forethought in privacy management.[19] By embedding privacy measures at the foundational level, GenAI platforms can eschew the pitfalls of retroactive amendments. For example, during a GenAI 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.[20]

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.[21] Drawing parallels, Apple's Siri epitomises this with its on-device personalisation, minimising data transfers to external servers and consequently bolstering user privacy.[22] 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.[23]

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.[24] This includes exhaustive Privacy Impact Assessments[25], the adoption of a privacy-centric

---

[19] 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

[20] 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

[21] 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/

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

[23] 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

[24] 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

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

architectural blueprint[26], immediate data anonymisation post-collection[27], and data acquisition limitations to essential elements only[28]. 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.[29] 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[30], 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,[31] 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.[32] This calls for meticulous field-level encryption of

---

[26] Fung, A., Yu, H., & Wright, J. (2021). Privacy-preserving artificial intelligence: A survey. *ACM Computing Surveys*, *54*(4), 1-37.

[27] 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.

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

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

[30] 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.

[31] 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

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

sensitive data[33] and rigorous role-based access control mechanisms during development[34]. Periodic external security audits, paired with transparent data retention and erasure protocols, ensure continued adherence to privacy norms.[35] 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.[36] 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.[37] Scholars like Solove have underscored that institutions collecting data should be transparent about their practices and held accountable for breaches and misuse.[38] 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 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.[39] Through such measures, GenAI platforms can underscore their unwavering allegiance to user-focused privacy in the

---

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

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

[35] 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.

[36] 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

[37] 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.

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

[39] Privacy in Context: Technology, Policy, and the Integrity of Social Life" (2010)

https://hci.stanford.edu/courses/cs047n/readings/Privacy_in_Context.pdf

design, development and deployment stage of the GenAI lifecycle fostering user empowerment.[40] In a notable instance, OpenAI, under the direction of the Italian data protection authority, introduced modifications to its privacy practices.[41] 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.[42] 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.[43] Such practices enable GenAI platforms to judiciously fulfil legal requirements while concurrently buttressing user trust.

ix. Data Minimisation

In embracing the principle of data minimisation,[44] GenAI platforms ought to champion a precision-focused approach to data collection[45]—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

---

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

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

[42] 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/

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

[44] 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

[45] 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.

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[46] necessitates a stringent confinement of data usage to its preordained intent.[47] 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.[48] 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.[49]

xi. Right to Erasure

The "Right to Erasure" is an indispensable tenet within GenAI platforms.[50] It magnifies user trust, bequeathing data sovereignty to individuals, thus potentially invigorating platform participation.[51] 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.[52] 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.

---

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

[47] 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.

[48] 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

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

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

[51] 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.

[52] 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

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.[53] 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.[54] For instance, a GenAI feedback tool might strip all user reviews of 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[55]. 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.[56]

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,

---

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

[54] 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.

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

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

offering guardians oversight over AI interactions with minors.[57] 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.[58] 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.[59] Measures to forestall the creation of deep fakes— manipulated visual 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.[60] While these platforms can generate unpredictable outputs, or hallucinate,[61] 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.[62] 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.[63] 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

---

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

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

[59] 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.

[60] 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

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

[62] 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.

[63] 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

feedback loops within an accountability framework allows for constant assessment 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.[64] 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[65]. 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.

7.5. 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 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.

References

[64] 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

[65] 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.

APPLE. How Siri Works. 27 jan. 2023. Available at: <https://support.apple.com/en-us/HT20704>.

ARTIFICIAL INTELLIGENCE/MACHINE LEARNING RISK & SECURITY WORKING GROUP (AIRS). Artificial Intelligence Risk & Governance. Wharton School of the University of Pennsylvania. Available at: <https://aiab.wharton.upenn.edu/research/artificial-intelligence-risk-governance/>.

BAROCAS, S.; HARDT, M.; NARAYANAN, A.; SELBST, A. D. The ethics of artificial intelligence: Mapping the debate. Nature, v. 569, n. 7755, p. 525-531, 2019.

BAROCAS, S.; NISSENBAUM, H. 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, 2014.

BAROCAS, S.; SELBST, A. D. The future of artificial intelligence and privacy. Harvard Law Review, v. 131, n. 1, p. 193-238, 2016.

BARREDO ARRIETA, A.; BOTTA, A.; DONINI, P.; IVKOVIC, M. Explainable AI: Concepts, taxonomies, opportunities and challenges towards responsible AI. Artificial Intelligence, v. 277, p. 1-35, 2020.

BARTH, A.; VOLKAMER, M.; SADEH, N. Privacy-preserving generative AI: A survey on mechanisms and challenges. ACM Computing Surveys, v. 53, n. 1, p. 1-38, 2020.

BARTOLI, A.; DE MATTEIS, S.; MAGGI, F. Privacy-preserving data retention and erasure in generative AI platforms: A survey. ACM Computing Surveys, v. 55, n. 1, p. 1-36, 2022.

BERRADA, M.; JACOVIDES, A.; OUEDRAOGO, A. Navigating intellectual property rights in the era of generative AI: The crucial role of educating judicial actors. UNESCO, fev. 2023. Available at: <https://www.unesco.org/en/articles/navigating-intellectual-property-rights-era-generative-ai-crucial-role-educating-judicial-actors?TSPD_101_R0=080713870fab2000186b2cce03d54213b1bdc1c26f57b6eea3d762f6edc2b3144e3067cff8>.

BHARGAVA, H. K.; KANTARCIOGLU, M.; ZHANG, L. Privacy-centric architectural blueprint for artificial intelligence systems. IEEE Security & Privacy, v. 18, n. 5, p. 68-75, 2020.

BYGRAVE, P. The principle of purpose limitation. In: DE HERT, P.; DE WEVER, P. (Eds.). Data protection law: A practical guide. Cambridge University Press, 2014, p. 33-48.

CARAHER, T. P.; ANDERSON, R. J. Socio-technical systems theory. In: Business. Leeds.ac.uk. Available at: <https://business.leeds.ac.uk/research-stc/doc/socio-technical-systems-theory#:~:text=Socio%2Dtechnical%20theory%20has%20at,parts%20of%20a%20complex%20systems>.

CAVOUKIAN, A. Privacy by design: The seven foundational principles. Information & Privacy Commissioner of Ontario, 2009.

CAVOUKIAN, A.; SINGH, S. The importance of purpose limitation in the age of Gen AI. Computer Law & Security Review, v. 36, n. 4, p. 1021-1034, 2020.

CHALLEN, K.; JONES, N. Risk management in the artificial intelligence act. European Journal of Risk Regulation, v. 4, n. 1, p. 25-42, 2023. Available at: <https://www.cambridge.org/core/journals/european-journal-of-risk-regulation/article/risk-management-in-the-artificial-intelligence-act/2E4D5707E65EFB3251A76E288BA74068>.

CHOUDHARY, S.; AGGARWAL, C. C. Differential privacy in healthcare: A survey. IEEE Transactions on Knowledge and Data Engineering, v. 29, n. 9, p. 2189-2210, 2017.

COHEN, J. E. Configuring the networked self: Law, code, and the play of everyday practice. Yale University Press, 2012.

CRAWFORD, K.; SCHULTZ, J. Big data and the child protection imperative. New Media & Society, v. 16, n. 1, p. 196-214, 2014.

DANAHER, J.; DEVLIN, K. Generative AI in healthcare. Nature Medicine, v. 26, n. 10, p. 1355-1357, 2020.

DANKS, D.; NIELSEN, M. Generative AI in finance. Journal of Financial Stability, v. 43, p. 100958, 2021.

DE SILVA, D.; ALAHAKOON, D. An artificial intelligence life cycle: From conception to production. Patterns, v. 3, n. 6, p. 100489, 2022.

DIETTERICH, T.; HOHMAN, M. Generative AI in manufacturing. Manufacturing Letters, v. 18, p. 1-5, 2019.

ERHARDT, J.; KRISHNAN, V. Designing generative AI to work for people with disabilities. Harvard Business Review, ago. 2023. Available at: <https://hbr.org/2023/08/designing-generative-ai-to-work-for-people-with-disabilities>.

ERLINGSSON, Ú.; KANTARCIOGLU, M.; ZHANG, L. Privacy-preserving personalization in voice assistants. ACM Transactions on Information Systems Security (TISSEC), v. 23, n. 3, p. 1-33, 2020.

EUROPEAN COMMISSION. AI and interpretability: Policy briefing. 2020. Available at: <https://ec.europa.eu/futurium/en/system/files/ged/ai-and-interpretability-policy-briefing_creative_commons.pdf>.

EUROPEAN COMMISSION. Regulatory framework for AI. Digital-Strategy.ec.europa.eu. Available at: <https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai>.

EUROPEAN PARLIAMENT. The European Union's Artificial Intelligence Act: A briefing. 2023. Available at: <https://www.europarl.europa.eu/RegData/etudes/BRIE/2023/739342/EPRS_BRI(2023)739342_EN.pdf>.

EUROPEAN UNION AGENCY FOR NETWORK AND INFORMATION SECURITY. Privacy notice. Available at: <https://gdpr.eu/privacy-notice/>.

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

EUROPEAN UNION. Regulation (EU) 2022/765 of the European Parliament and of the Council on artificial intelligence (Artificial Intelligence Act). Official Journal of the European Union, 2022.

FUNG, A.; YU, H.; WRIGHT, J. Privacy-preserving artificial intelligence: A survey. ACM Computing Surveys, v. 54, n. 4, p. 1-37, 2021.

GARANTE PER LA PROTEZIONE DEI DATI PERSONALI. Available at: <https://www.garanteprivacy.it/web/garante-privacy-en/home_en>.

GAUDIOSI, J. ChatGPT is once again available in Italy after a temporary ban. Engadget, 1 abr. 2023. Available at: <https://www.engadget.com/chatgpt-is-once-again-available-in-italy-after-a-temporary-ban-195716663.html>.

GENERAL DATA PROTECTION REGULATION. Available at: <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32016R0679>.

GHOSH, A.; KANTARCIOGLU, M. Differential privacy in generative AI: A survey. ACM Computing Surveys, v. 53, n. 1, p. 1-38, 2020. See also: DWORK, C. Differential privacy. Automated Decision Making, v. 1, n. 2, p. 23-40, 2006.

GRAND VIEW RESEARCH. The rise of generative AI platforms: A market research report. Grand View Research, 2023.

GREENBERG, M. ChatGPT users can now ask OpenAI to delete their data. TechCrunch, 2 maio 2023. Available at: <https://techcrunch.com/2023/05/02/chatgpt-delete-data/?guccounter=1>.

GUPTA, A.; DAS, A. K. A survey on data anonymization techniques for general-purpose artificial intelligence systems. ACM Computing Surveys, v. 55, n. 2, p. 1-41, 2022.

GUPTA, A.; KAGAL, L. Preventing deanonymization in GenAI platforms: A survey of techniques and challenges. ACM Computing Surveys, v. 53, n. 2, p. 1-35, 2020.

HAAKMAN, M.; CRUZ, L.; HUIJGENS, H.; VAN DEURSEN, A. AI lifecycle models need to be revised: An exploratory study in Fintech. Empirical Software Engineering, v. 26, n. 5, p. 1-29, 2021.

KANTARCIOGLU, M.; DASGUPTA, K. The right to erasure in the age of artificial intelligence: Challenges and opportunities. IEEE Security & Privacy, v. 17, n. 4, p. 78-85, 2019.

LI, X.; ZHANG, L.; KANTARCIOGLU, M.; CHOO, K.-K. R. Field-level encryption: A survey. ACM Computing Surveys, v. 49, n. 4, p. 1-35, 2017.

MARTENS, M.; RAU, M.; SCHERRER, S. Data minimization in the age of big data: A review of concepts, methods, and tools. Computer Law & Security Review, v. 34, n. 1, p. 107-124, 2018.

MICHALSONS. Privacy impact assessments for generative AI. Michalsons, 15 fev. 2023. Available at: <https://www.michalsons.com/blog/privacy-impact-assessments-for-generative-ai/65772>.

MONTREAL DECLARATION FOR RESPONSIBLE DEVELOPMENT OF ARTIFICIAL INTELLIGENCE. 2018. Available at: <https://recherche.umontreal.ca/english/strategic-initiatives/montreal-declaration-for-a-responsible-ai/>.

MULLIGAN, D. K.; KING, J. L. Bridging the gap between privacy and design. University of Pennsylvania Law Review, v. 159, n. 5, p. 1087-1174, 2011.

OPENAI. ChatGPT: Generative pre-trained transformer for conversational applications. 23 fev. 2022. Available at: <https://openai.com/blog/chatgpt/>.

OPENAI. GPT-4 Technical Report. 2023. Available at: <https://cdn.openai.com/papers/gpt-4.pdf>.

ORGANISATION FOR ECONOMIC CO-OPERATION AND DEVELOPMENT (OECD). Classification of artificial intelligence: A two-pager. 2022. Available at: <https://wp.oecd.ai/app/uploads/2022/02/Classification-2-pager-1.pdf>.

PRIVACY IN CONTEXT: Technology, Policy, and the Integrity of Social Life. 2010. Available at: <https://hci.stanford.edu/courses/cs047n/readings/Privacy_in_Context.pdf>.

SOLOVE, D. J. Understanding privacy. Harvard Law Review, v. 125, n. 3, p. 421-549, 2008.

SUSARLA, A.; CHUI, M.; OSBORNE, M. The black box of AI: How to mitigate the risks of unexplained bias. Harvard Business Review, mar. 2020.

THE ARTIFICIAL INTELLIGENCE ACT: A new regulation for artificial intelligence in the European Union. Available at: <https://www.europarl.europa.eu/RegData/etudes/STUD/2020/641530/EPRS_STU(2020)641530_EN.pdf>.

THE FUTURE OF LIFE INSTITUTE. Asilomar AI Principles. 2017. Available at: <https://futureoflife.org/ai-principles/>.

THE GUARDIAN. I didn't give permission: Do AI's backers care about data law breaches? 10 abr. 2023. Available at: <https://www.theguardian.com/technology/2023/apr/10/i-didnt-give-permission-do-ais-backers-care-about-data-law-breaches>.

UNIVERSITY OF MELBOURNE. Advice for students regarding Turnitin and AI writing detection. 8 mar. 2023. Available at: <https://academicintegrity.unimelb.edu.au/plagiarism-and-collusion/artificial-intelligence-tools-and-technologies/advice-for-students-regarding-turnitin-and-ai-writing-detection>.

VEALE, M.; BRASS, A. The stakeholders of artificial intelligence: A literature review. Ethics and Information Technology, v. 22, n. 1, p. 1-18, 2020.

VEALE, M.; WALSH, M. The copyright challenges of generative AI. AI & Society, v. 37, n. 1, p. 1-18, 2022. See also: CALO, R.; BUCCAFUSCO, C. Generative AI and the future of copyright. The Yale Law Journal, v. 130, n. 1, p. 1-60, 2020.

WALSH, M.; VEALE, M. Generative AI in art and design. AI & Society, v. 37, n. 1, p. 1-18, 2022.

WARDRIP-FRUIN, N.; MATEAS, M. Generative AI in media and entertainment. ACM Transactions on Computer-Human Interaction (TOCHI), v. 25, n. 6, p. 1-27, 2018.

WORLD ECONOMIC FORUM. The Future of Jobs: Jobs and Skills in 2030. 2020. Available at: <https://www.weforum.org/reports/the-future-of-jobs-report-2020/>.

ZHANG, L.; BARTH, A.; VOLKAMER, M. Transparency reports for generative AI platforms: A review and research agenda. ACM Computing Surveys, v. 55, n. 1, p. 1-36, 2022.

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, Senior 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.

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 behaviour. 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.

8.1. 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.

8.2. Landscape study of several frameworks

8.2.1. 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.

8.2.2. 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.

8.2.3. 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.

8.2.4. 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.

8.3. 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 centred 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.

8.4. 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.

8.5. 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[1]. Building upon transparency, Explainable AI (XAI)[2] 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[3]. A clear understanding of the decision-making processes enables researchers to

[1] 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

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

[3] 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/

validate AI systems rigorously and identify potential biases or errors that may arise during model operation.

8.5.1. 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)[4] and SHAP (Shapley Additive Explanations)[5] 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 behaviour.

In addition to model interpretability techniques, user-friendly dashboards play a crucial role in enhancing transparency for end-users[6]. 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[7] and maintaining detailed documentation of AI model development. Data lineage enables AI developers and technical auditors to trace the origin and transformation of data 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[8]. This

---

[4] 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/

[5] 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/

[6] 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

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

[8] 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/

documentation promotes transparency by enabling other researchers and auditors to validate the AI system's performance and scrutinize its decision-making processes.

8.5.2. 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[9]. 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[10]. 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.

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[11]. 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.

8.6. Accountability

Accountability is a critical principle that underpins the entire lifecycle of an AI system[12]. 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

---

[9] 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

[10] 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/

[11] 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

[12] 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

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[13].

8.6.1. 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 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[14].

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.[15] 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[16], 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.

---

[13] 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

[14] 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

[15] 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

[16] https://arxiv.org/pdf/2001.00973.pdf

8.6.2. Non-technical Level

At the non-technical level, encouraging AI developers and users to adopt ethical guidelines and industry best practices is instrumental in fostering accountability[17]. 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[18]. 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[19]. 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 instil 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[20]. Compliance with these standards ensures that AI systems undergo rigorous scrutiny and assessment to meet predefined criteria for fairness, explainability, and safety.

8.7. Fairness and Non-discrimination

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

---

[17] 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

[18] 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

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

[20] 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

systematic unfairness and discrimination in AI decisions.[21] 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).

8.7.1. 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.[22] 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.[23] 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.

8.7.2. 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.[24] This interdisciplinary approach ensures that AI systems are designed and deployed in a manner that considers ethical and societal considerations.

---

[21] 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

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

[23] 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

[24] 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

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.[25] Increased awareness can also empower individuals to demand fair and transparent AI solutions.

8.8. Reliability and Safety/Robustness

Reliability and safety/robustness are fundamental principles in ensuring the trustworthy operation of AI systems.[26] 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 inputs gracefully.[27] 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.

8.8.1. 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.[28] Through adversarial training, the AI system is trained to recognize and defend against these adversarial inputs, making it more resilient to potential attacks.

---

[25] 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

[26] 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.

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

[28] 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

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.[29] 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 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.

8.8.2. 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.[30] 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.[31] 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.

8.9. 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.[32] When users entrust their data to AI systems, they expect

---

[29] 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

[30] 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

[31] 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

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

that their personal information will be handled with utmost care and 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.

8.9.1. Technical Level

To ensure robust privacy protection in AI systems, developers and organizations must implement a range of data privacy measures.[33] 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.[34] 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.[35] 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, 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.[36] Providing clear and easily understandable explanations empowers users to make informed decisions about sharing their data, giving them greater control over their information.

---

[33] 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.

[34] 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.

[35] 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

[36] 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

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.[37] 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.

8.9.2. 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.[38] 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 instil confidence in users, making them more likely to choose AI systems from certified organizations over others.[39]

Ethical considerations also come into play, as organizations can adopt ethical frameworks that prioritize user consent, fairness, and responsible data handling.[40] 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.

8.10. 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 behaviour, 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

---

[37] 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

[38] 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

[39] 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

[40] 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

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 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.

References

ABDAR, M. et al. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. Information Fusion, v. 76, p. 243–297, 2021. Available at: <https://doi.org/10.1016/j.inffus.2021.05.008>.

ALI, S. et al. Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence. Information Fusion, v. 99, p. 101805, 2023. Available at: <https://doi.org/10.1016/j.inffus.2023.101805>.

ANDREOTTA, A.; KIRKHAM, N.; RIZZI, M. AI, big data, and the future of consent. AI & Society, v. 37, n. 4, p. 1715–1728, 2021. Available at: <https://doi.org/10.1007/s00146-021-01262-5>.

BIAS AND ETHICAL CONCERNS IN MACHINE LEARNING. Available at: <https://www.isaca.org/resources/isaca-journal/issues/2022/volume-4/bias-and-ethical-concerns-in-machine-learning>.

BUILDING TRANSPARENCY INTO AI PROJECTS. Harvard Business Review, 20 jun. 2022. Available at: <https://hbr.org/2022/06/building-transparency-into-ai-projects>. Acesso em: 25 ago. 2023.

ENDSLEY, M. R. Supporting Human-AI Teams: Transparency, explainability, and situation awareness. Computers in Human Behavior, v. 140, p. 107574, 2023. Available at: <https://doi.org/10.1016/j.chb.2022.107574>.

GOYAL, S. et al. A survey of Adversarial Defenses and Robustness in NLP. ACM Computing Surveys, v. 55, n. 14s, p. 1–39, 2023. Available at: <https://doi.org/10.1145/3593042>.

HUANG, A.; LI, J.; SHANKAR, N. Interpretability – Machine Learning Blog | ML@CMU. Carnegie Mellon University, 31 ago. 2020. Available at: <https://blog.ml.cmu.edu/2020/08/31/6-interpretability/>.

HUANG, A.; LI, J.; SHANKAR, N. Interpretability – Machine Learning Blog | ML@CMU. Carnegie Mellon University, 31 ago. 2020. Available at: <https://blog.ml.cmu.edu/2020/08/31/6-interpretability/>.

HUNKENSCHROER, A. L.; KRIEBITZ, A. Is AI recruiting (un)ethical? A human rights perspective on the use of AI for hiring. AI And Ethics, v. 3, n. 1, p. 199–213, 2022. Available at: <https://doi.org/10.1007/s43681-022-00166-4>.

IEEE SA. The Ethics Certification Program for Autonomous and Intelligent Systems (ECPAIS). IEEE Standards Association, 31 jul. 2023. Available at: <https://standards.ieee.org/industry-connections/ecpais/>.

INTERPRETML: A toolkit for understanding machine learning models. Microsoft, 18 maio 2020. Available at: <https://www.microsoft.com/en-us/research/uploads/prod/2020/05/InterpretML-Whitepaper.pdf>. Acesso em: 25 ago. 2023.

JIN, D. et al. A survey on fairness-aware recommender systems. Information Fusion, v. 100, p. 101906, 2023. Available at: <https://doi.org/10.1016/j.inffus.2023.101906>.

KONIGSTORFER, F.; THALMANN, S. AI Documentation: A path to accountability. Journal of Responsible Technology, v. 11, 2022, p. 100043. Available at: <https://www.sciencedirect.com/science/article/pii/S2666659622000208/>.

LAWTON, G. AI transparency: What is it and why do we need it? CIO, 2023. Available at: <https://www.techtarget.com/searchcio/tip/AI-transparency-What-is-it-and-why-do-we-need-it>.

LAWTON, G. The future of trust will be built on data transparency. CIO, 2021. Available at: <https://www.techtarget.com/searchcio/feature/The-future-of-trust-must-be-built-on-data-transparency>.

MANTELERO, A. The social and ethical component in AI systems design and management. In: Information technology & law series. p. 93–137, 2022. Available at: <https://doi.org/10.1007/978-94-6265-531-7_3>.

MICROSOFT. Responsible AI Standard, v2: General Requirements. Available at: <https://blogs.microsoft.com/wp-content/uploads/prod/sites/5/2022/06/Microsoft-Responsible-AI-Standard-v2-General-Requirements-3.pdf>. Acesso em: 25 ago. 2023.

MOREY, T. Customer Data: Designing for transparency and trust. Harvard Business Review, 1 set. 2020. Available at: <https://hbr.org/2015/05/customer-data-designing-for-transparency-and-trust>.

MOREY, T. Customer Data: Designing for transparency and trust. Harvard Business Review, 1 set. 2020. Available at: <https://hbr.org/2015/05/customer-data-designing-for-transparency-and-trust>.

MSTELLER-AI. Responsible and trusted AI - Cloud Adoption Framework. 28 jul. 2023. Available at: <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>.

NOVELLI, C.; TADDEO, M.; FLORIDI, L. Accountability in artificial intelligence: what it is and how it works. AI & Society, 2023. Available at: <https://doi.org/10.1007/s00146-023-01635-y>.

QUACH, S. et al. Digital technologies: tensions in privacy and data. Journal of the Academy of Marketing Science, v. 50, n. 6, p. 1299–1323, 2022. Available at: <https://doi.org/10.1007/s11747-022-00845-y>.

RAJI, I. D. et al. Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing. arXiv, 3 jan. 2020. Available at: <http://arxiv.org/abs/2001.00973>. Acesso em: 29 jul. 2024.

REINHARDT, K. Trust and trustworthiness in AI ethics. AI And Ethics, 2022. Available at: <https://doi.org/10.1007/s43681-022-00200-5>.

RESPONSIBLE AI | AI Ethics & Governance. Accenture. Available at: <https://www.accenture.com/in-en/services/applied-intelligence/ai-ethics-governance>.

RYAN, M.; STAHL, B. C. Artificial intelligence ethics guidelines for developers and users: clarifying their content and normative implications. Journal of Information, Communication and Ethics in Society, v. 19, n. 1, p. 61–86, 2020. Available at: <https://doi.org/10.1108/jices-12-2019-0138>.

SANFORD, S. How to Build Accountability into Your AI. Harvard Business Review, 30 ago. 2021. Available at: <https://hbr.org/2021/08/how-to-build-accountability-into-your-ai>.

SINGH, R. Trustworthy AI. arXiv, 2 nov. 2020. Available at: <https://arxiv.org/abs/2011.02272>.

TACKLING BIAS IN ARTIFICIAL INTELLIGENCE (AND IN HUMANS). McKinsey, 6 jun. 2019. Available at: <https://www.mckinsey.com/featured-insights/artificial-intelligence/tackling-bias-in-artificial-intelligence-and-in-humans>.

THEOTO, T. et al. Responsible AI for an era of tighter regulations. BCG Global, 28 jul. 2023. Available at: <https://www.bcg.com/publications/2022/acting-responsibly-in-tight-ai-regulation-era>.

UNESCO. Recommendation on the ethics of artificial intelligence. Available at: <https://en.unesco.org/about-us/legal-affairs/recommendation-ethics-artificial-intelligence>.

VAN RIJMENAM, C. Privacy in the age of AI: Risks, challenges and solutions. Dr Mark Van Rijmenam, CSP | Strategic Futurist Speaker, 2023. Available at: <https://www.thedigitalspeaker.com/privacy-age-ai-risks-challenges-

solutions/#:~:text=Organisations%20that%20use%20AI%20must,whose%20data%20has%20been%20compromised>.

VORRAS, A.; MITROU, L. 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. Available at: <https://link.springer.com/chapter/10.1007/978-3-030-69583-5_10/>.

WHAT IS DATA ANONYMIZATION | Techniques, Pros, Cons, and Use Cases. K2View. Available at: <https://www.k2view.com/what-is-data-anonymization/#:~:text=Data%20anonymization%20transforms%20personal%20or,privacy%20laws%20and%20heighten%20security>.

WHAT IS DATA LINEAGE? Informatica. Available at: <https://www.informatica.com/resources/articles/what-is-data-lineage.html>. Acesso em: 25 ago. 2023.

WHAT IS EXPLAINABLE AI? IBM. Available at: <https://www.ibm.com/watson/explainable-ai>. Acesso em: 25 ago. 2023.

WILLIAMS, R. et al. From transparency to accountability of intelligent systems: Moving beyond aspirations. Data & Policy, v. 4, 2022. Available at: <https://doi.org/10.1017/dap.2021.37>.

WILLIAMS, R. et al. From transparency to accountability of intelligent systems: Moving beyond aspirations. Data & Policy, v. 4, 2022. Available at: <https://doi.org/10.1017/dap.2021.37>.

XU, J. Algorithmic Solutions to Algorithmic Bias: A Technical Guide. Medium, 10 dez. 2021. Available at: <https://towardsdatascience.com>.

PART 3:
WESTERN PERSPECTIVES ON AI GOVERNANCE

Pre-print version of Belli L. Gaspar. W.B. The Quest for AI Sovereignty, Transparency and Accountability. Springer-Nature. (2025).

9. AI and EU: A 'third way'?

Yves Poullet, Emeritus Professor and rector at the University of Namur (Belgium), IFAP/UNESCO Vice chairman, Member of the Belgian Data Protection Authority and of the CCF/Interpol, Member of the Royal Academy of Belgium.

Abstract[1,2]

On 8 December, the European trilogue (EU Council of Ministers, EU Commission and EU Parliament) agreed on the main lines of the European regulation of artificial intelligence. This regulation is the culmination of the European strategy on the development of our information society and is justified by the desire for a sovereign Europe both in the protection of its values and its economic interests. The contribution aims to show how the AI regulation illustrates a new regulatory approach, already present in the other texts but not in such a way. After these remarks, the contribution intends to analyse at the light of the final compromise but also by referring to the previous versions of the text the main provisions of the AI Act. How has the text finally integrated the 'foundation models'? How has the text enlarged the risks to consider by the main actors and following which values they have to assess and if needed to mitigate these risks? Which categories of AI systems, the AI Act is distinguishing according to the gravity of the risks generated and which obligations are linked to each of these categories? Which governance including as regards the standardisation procedures has been put into place to control the respect of these obligations and the future developments of the AI systems? Finally, how the text is balancing the principle of precaution with the concern of innovation?

Introduction

The European Union is restructuring its regulatory policy developed about evolving technologies, 3D, Blockchain, Internet of Things and AI, including AI generative models and applications. These technologies, more and more, are ubiquitous and determine our way of life, our economic development, and our administrations' public services[3] through certain actors, in particular the platforms – some of which are called "Gatekeepers" in the EU[4] – of

---

[1] This text has been presented at 2023 APCA-ANPOR Annual Conference – organized by Tsinghua University, Beijing, November 5th, 2023.

[2] The text on EU AI regulation established by the COREPER (Representatives of the EU Member States) has been delivered the 2nd of February 2024. It does represent the translation of the compromise between EU Parliament, EU Council of Ministers and EU Commission obtained the 9th of December 2023 and even if formal approvals by the Council and the Parliament are still necessary, the COREPER text might be considered as the final one. Our reflections are based on the compromise, are based on the EU Council of Ministers and EURACTIV reports. We might not ensure that no future modifications have been introduced during the last step. Since the date of writing, a final version of the EU AI Act has been adopted by the European Parliament. This final version could not be considered in time for the closing of this volume. However, the editors and author still consider the reflections contained herein relevant for analysis.

[3] About that 'revolution' read our developments and the numerous references, Le RGPD face aux défis de l'intelligence artificielle, Coll. Cahiers du CRIDS n° 48, Larcier, Bruxelles, 2020, p. 15 and ff.

[4] According to the Cambridge Dictionary, the term "gatekeeper", explicitly used by the Digital Market Act (DMA), refers to "a person whose job Is to open and close a gate and to prevent people entering without permission; someone who has the power to decide who gets particular resources and opportunities, and who does not". The notion of controlling public access to a resource or service is therefore central. It is easily applied

our information society (social networks and search engines). In recent times, the European Union has proactively multiplied the regulatory texts relating to various aspects of the digitalization of society on both issues: the technologies and the actors. The number of acts and drafts is impressive and grants practically no margin of manoeuvre to the member-states. A non-exhaustive list of such recent policymaking efforts includes the Data Governance Act (DGA)[5], the Digital Market Act (DMA)[6], the Digital Service Act (DSA)[7], the Data Act[8], the EU Cybersecurity Act[9], the Media Freedom Act[10], and the AI Act[11]. All have been enacted,

---

in the digital world, when both economic explanations (large economies of scale and significant network effects) and/or technical explanations (imposition of a proprietary standard such as Apple's for App stores on iOS devices) are combined to explain why some companies have these access controls. At this moment, 19 companies have been designated by the EU Commission as "Gatekeepers".

[5] Regulation (EU) 2022/868 of the European Parliament and of the Council of 30 May 2022 on European data governance and amending Regulation (EU) 2018/1724 (Data Governance Act, PE/85/2021/REV/1, **OJ L 152, 3.6.2022, p. 1–44. The DGA main aim is to facilitate and organize, to authorize the use of their data by public bodies in the context of specific public needs. 'Data sharing' in the EU context might be considered as a condition for the building-up of big data needed for AI development.**

[6] Regulation (EU) 2022/1925 of the European Parliament and of the Council of 14 September 2022 on contestable and fair markets in the digital sector and amending Directives (EU) 2019/1937 and (EU) 2020/1828 (Digital Markets Act),PE/17/2022/REV/1, **OJ L 265, 12.10.2022, p. 1–66: 'The Digital Markets Act (DMA) will ban certain practices used by large platforms acting as "gatekeepers" and enable the Commission to carry out market investigations and sanction non-compliant behaviour.'**

[7] Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market For Digital Services and amending Directive 2000/31/EC (Digital Services Act), PE/30/2022/REV/1, **OJ L 277, 27.10.2022, p. 1–102. '***The DSA regulates online intermediaries and platforms such as marketplaces, social networks, content-sharing platforms, app stores, and online travel and accommodation platforms. Its main goal is to prevent illegal and harmful activities online and the spread of disinformation. It ensures user safety, protects fundamental rights, and creates a fair and open online platform environment.*'

[8] REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL on harmonised rules on fair access to and use of data (Data Act), Brussels, 15 November 2023, 2022/0047 (COD) PE-CONS 49/23'. '*The Data Act aims to maximise the value of data in the economy by ensuring that a wider range of stakeholders gain control over their data and that more data is available for innovative use, while preserving incentives to invest in data generation.'*

[9] Regulation (EU) 2019/881 of the European Parliament and of the Council of 17 April 2019 on ENISA (the European Union Agency for Cybersecurity) and on information and communications technology cybersecurity certification and repealing Regulation (EU) No 526/2013 (Cybersecurity Act), PE/86/2018/REV/1, **OJ L 151, 7.6.2019, p. 15–69:** '*The **EU Cybersecurity Act** introduces an EU-wide cybersecurity certification framework for ICT products, services and processes. Companies doing business in the EU will benefit from having to certify their ICT products, processes and services only once and see their certificates recognised across the European Union.'*

[10] Proposal for a REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL establishing a common framework for media services in the internal market (European Media Freedom Act) and amending Directive 2010/13/EU, Brussels, 16.9.2022, COM(2022) 457 final : *'... The European Union remains a stronghold for free media, setting a standard as a democratic continent. Yet, there are increasingly worrying trends. Building on past efforts, the Commission has taken a number of measures to protect media freedom and pluralism in the EU'.*

[11]Proposal for a regulation laying down harmonised rules on artificial intelligence, 21 April 2021, COM (2021)206 Final 2021/0106. See on that text, the Council of the European Union amendments, Brussels, 25 November 2022 (OR. en) 14954/22 and the 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 harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts (COM (2021)0206 – C9-0146/2021 –2021/0106(COD) and the final compromise reached by the EU authorities (EU

except the Media Freedom Act, and constitute all together the enactment of the EU Data Strategy.

The texts so listed consider, first, the deep modifications of the digital market, the merging of the telecommunications, audiovisual and information society services, second the ubiquitous presence of certain actors: the GAFAM[12] or the BATX[13], and third, the increasing impact of our digital society not only on our way of doing business or conducting public affairs but also on our life, identity, and liberties. Through these texts, the Union's desire to chart a "Third Way" forward in terms of the development of our digital society which must be "human centred," distinct from that of the United States, and China, and based in particular on respect for human rights and the pre-eminence of the human, following the famous formula: *"Human in the loop, Human on the loop and Human in command"*[14].

Beyond the multiplication of these regulatory texts, it is interesting to highlight a certain number of the characteristics of this EU regulatory approach[15]: how the EU authorities have imposed a coregulatory model instead of self-regulation[16] and how they are achieving a full

---

Commission, EU Parliament, EU Council of Ministers). December, the 8th 2023. On that compromise, read the report of the EU Council of Ministers available at: https://www.consilium.europa.eu/en/press/press-releases/2023/12/09/artificial-intelligence-act-council-and-parliament-strike-a-deal-on-the-first-worldwide-rules-for-ai/. See also https://www.iapp.org/news/a/eu-institutions-work-toward-final-ai-act-compromises/#:~:text=Euractiv%20reports. As underlined above, **the final text will be delivered in the begin of 2024 according to the compromise. Our reflections on the compromise are based on the EU Council of Ministers and EURACTIV reports.**

[12] Abbreviation still used for the 5 major US bigtech companies (2010): Google, Apple, Facebook (now Meta), Amazon and Microsoft

[13] Abbreviation still used for the designating the four major Chinese bigtech companies (2010): Baidu, Alibaba, Tencent and Xiaomi.

[14] This formula has been asserted by the EU HLGE (High level Group of experts) in charge to analyse the ethical problems linked with the development of the AI technologies. The Group published the 19th of April its "Ethics guidelines for trustworthy AI" (https://digital-strategy.ec.europa.eu/en/library/ethics-guide lines-trustworthy-ai)

[15] About the main characteristics of the new regulatory approach present in the different regulatory texts constituting the legal framework of our digital society, our contribution: "Towards a New EU Regulatory Approach of the Digital Society, ERDAL, 2022, Vol. 1, p. 113 and ff.

[16] On the distinction between the three concepts: Self-regulation, - Co-regulation and Pubic regulation, read: Resolution of 9 September 2010 on Better Lawmaking 14 (P7TA(2010)0311). In this Resolution, the European Parliament "*warns against abandoning necessary legislation in favour of self-regulation or co-regulation or any other non-legislative measure; (it) believes that the consequences of such choices should be subject to careful examination in each case, in accordance with Treaty law and the roles of the individual institutions*". It also "*stresses, at the same time, that soft law should be applied with the greatest of care and on a duly justified basis, without undermining legal certainty and the clarity of existing legislation, and after consultation of Parliament as underlined in its resolution on a revised Framework Agreement*". Points 17 and 22 of the Inter-institutional Agreement define both co-regulation and self-regulation: As regards self-regulation, point 22 stipulates: "Self-*regulation is defined as the possibility for economic operators, the social partners, non-governmental organisations or associations to adopt amongst themselves and for themselves common guidelines at European level (particularly codes of practice or sectoral agreements). As a general rule, this type of voluntary initiative does not imply that the Institutions have adopted any particular stance, in particular where such initiatives are undertaken in areas which are not covered by the Treaties or in which the Union has not hitherto legislated. As one of its responsibilities, the Commission will scrutinise self-regulation practices in order to verify that they comply with the provisions of the EC Treaty*.".

consistent EU market. Furthermore, EU recent regulations adopt an asymmetrical approach to regulate especially the major actors and in order to ensure the proportionality of their intervention and the effectiveness of their regulations legal control[17].The EU authorities promote a risk-based approach and of preventive measures, including the creation of internal compliance bodies, in addition to or instead of the traditional a posteriori legal control[18]. Section 1 of this chapter endeavours to analyse these different characteristics of the EU approach.

It would be impossible in the context of this brief contribution to develop each of these EU legislative initiatives as listed above. Section 2 will provide an overview of the EU policy through the AI Act taken as an example particularly significative of the EU approach.

The European conception of co-regulation does envisage this mechanism not as a way to prepare future public regulation[16] but as a tool for refining the content of the regulation enacted by the public bodies and for implementing concretely it. By doing that, the Agreement underlines the essential place of the co-regulation.: "*Co-regulation means the mechanisms whereby a community legislative act entrusts the attainment of the objectives defined by the legislative authority to parties which are recognized in the field (such as economic operators, the social partners, non-governmental organisations, or associations)". This definition induces a clear partition of the responsibilities of the State, from one part, and the private sector and other interested parties, from the other part, in the regulatory process: the legislative authorities have to fix the essential public policy objectives, when the means, by which they are met, are fixed together by the public and the private sectors. The private sector is mainly responsible for defining apart from the end result and objectives fixed by the legislative instruments, shortly to answer to the question: "How to implement them?*" (Y. Poullet, Governance Challenges: First Lessons from the WSIS ¿ An Ethical and Social Perspective, In: Innovation, Legitimacy, Ethics and Democracy, in Liber Amicorum J. Berleur, PUN, Namur, 2007). See also, Y. Poullet**,** « How to regulate the Internet : New paradigms for Internet Governance ?, in *E-Commerce Law and practice in Europe,* I WALDEN et J. HÖRNLE (ed.), Cambridge, Woodehead, Section 1, Chap. 2. More recently, the Study launched by the European Social and Economic Committee (EESC) : „Self-and Co-regulation": https:// www.eesc.europa.eu/sites/default/files/resources/docs/auto_coregulation_en--2.pdf.

[17] Another characteristic emerges in the most recent European Union texts, namely **asymmetrical regulation** of both the players and the applications operated, or products or services offered by them, depending on the risks (*risk-based approach*) associated with these actors, applications, products or services. In all the cases, this regulatory asymmetry is justified by the principle of proportionality, affirmed by Article 5(4) of the Treaty on European Union, which stipulates that the Union must not in exercising its powers do more than is necessary to achieve its objectives. To take examples, some European regulations impose heavier obligations on certain categories of actors. As regards the first category, provisions are imposed on communication and information platforms, such as the equal and transparent treatment of professional users by these necessary intermediaries. Similarly, the DSA imposed, only on '*very large platforms' (i*.e. those with a customer base equal to or greater than 10% of the European population), certain obligations, namely that to monitor content and audit recommendation systems. At the contrary, small and medium enterprises will benefit of the exemption of different provisions imposed by the AI Act or by other regulation in order to scrutinize the flow of information accessible through their services in order to fight against hate speech, terrorism, copyright infringement, …

[18] The **genuine risk-based approach** leads to the creation of new obligations when certain criteria proposed by the regulation indicate that higher risks are present. This approach is already used, but in a very limited way, in the provisions of the RGPD. Article 35 reserves the obligation to carry out an impact assessment only to processing operations presenting a "high risk" to the rights and freedoms of natural persons. The 2017 Regulation on medical devices (Regulation (EU) 2017/745 of the European Parliament and of the Council of 5 April 2017 on medical devices, **OJ L 117, 5.5.2017, p. 1–175)** similarly distinguishes between different classes of products and services according to the purpose of their use and the risks related to health and safety, and subjects "high risk" classes of products to conformity assessment procedures. The same idea runs through *the AI ACT*. The initial proposal sets out the prohibition of illegal practices of artificial intelligence (Art. 5); it establishes a system of control and management of high-risk AI systems (Art. 6.2). The DSA imposes additional obligations to the very large Platforms due to the fact they are creating "systemic risks" including risks for our democracy.

9.1. The EU response to AI technologies and its "Strategy for Data": from the 'third way' to EU sovereignty

What specific regulatory response is Europe providing to the challenges of those new digital technologies invading our lives, in particular with generative AI applications which afford to each of us the possibility of benefitting from their potential applications? Doesn't digital technology now stick to us, both figuratively and in reality? Does it not guide, for better or worse, our lives as well as those of companies and administrations? It is therefore important, and it is the role of the public authority, to map out the uses of a tool, which, increasingly, is the backbone of our economy, our society, our relationships, and ourselves. The introduction mentioned the European Union will lead a 'third way'. What is it about? The strategy is explicitly stated in the so-called White Paper[19] "A European Strategy for Data" from February 19, pronounced by the EU Commission's chairwoman, Ursula van der Leyen. Its implementation has since been carried out through texts that follow one another at an accelerated rhythm and go far beyond the sole issue of artificial intelligence.

As it will be emphasised, it is a regulatory policy on data, their creation, use, transmission, and impacts that Europe intends to develop in a coherent manner. This is indeed a "third way"[20] insofar as the European Union intends to conduct its own digital development policy. This EU policy is based on principles different than those of the two major competitors, the US and China. On one hand, the American policy[21] can, no doubt partly wrongly, be summarised as 'all for the market' and, more correctly, by the desire to maintain and develop the digital economy and the desire to maintain and develop American leadership. On the other hand, the Chinese policy is marked - but we are probably close to a caricature - by a strong State interventionism since AI must be at the service of the socio-economic governance by the State and the public security to the detriment of the individual freedoms of citizens[22].

---

[19] Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of The Regions, *A European strategy for data*, COM/2020/66, Brussels, 19 February 2020, final: "*The European data strategy aims to make the EU a leader in a data-driven society. Creating a single market for data will allow it to flow freely within the EU and cross sectors for the benefit of businesses, researchers, and public administrations. People, businesses, and organizations should be empowered to make better decisions based on insights from non-personal data, which should be available to all*".

[20] See European Commission, USA-China-EU plans for AI: where do we stand? January 2018, text available at: https:// monitor-industrial-ecosystems.ec.europa.eu/reports/other-reports/usa-china-eu-plans-ai-where-do-we-stand.

[21] The American model is founded on a coregulatory approach resulting from a dialog between the governmental authority and the private sector, especially the digital platforms as it is clearly asserted by « The Blueprint for an AI Bill of Rights : Making Automated Systems Work for the American People », published by the White House Office of Science and Technology Policy, October 2022; The recent "President Biden Issues Executive Order on Safe, Secure, and Trustworthy Artificial Intelligence" from October, 30, 2023 intends to be more imperative even it expressly refers to the previous t Blueprint. As said by the introduction: "*The Executive Order establishes new standards for AI safety and security, protects Americans' privacy, advances equity and civil rights, stands up for consumers and workers, promotes innovation and competition, advances American leadership around the world, and more. As part of the Biden-Harris Administration's comprehensive strategy for responsible innovation, the Executive Order builds on previous actions the President has taken, including work that led to voluntary commitments from 15 leading companies to drive safe, secure, and trustworthy development of AI.*

[22] See, recently the « interim Administrative Measures for Generative Artificial Intelligence Services » adopted at the 12th meeting of the Chinese Cyberspace Administration, The 23rd of May and approved by diverse Ministers and Commission are in force since the 15th of Augustus.

Definitely this EU strategy is founded on what the EU calls its 'digital sovereignty', defined in a positive sense as the possibility to assert its autonomy in its choice of society and therefore of the technology able to achieve the EU values[23]. Digital sovereignty has also another goal. Through that concept, Europe intends to eliminate intra-European barriers to the deployment of AI and, more generally, digital technology. The clearly stated ambition is economic: It means to enable the European Union "to compete[24] with the massive investments made by the competitors already mentioned, the *United States*[25] and *China.*"[2627], by imposing on its market its proper standards and its regulation including towards third

[23] The notion of 'technological' or 'digital sovereignty' has recently emerged as a means of promoting the notion of European leadership and strategic autonomy in the digital field. Strong concerns have been raised over the economic and social influence of non-EU technology companies, which threatens EU citizens' control over their personal data, and constrains both the growth of EU high technology companies and the ability of national and EU rule-makers to enforce their laws. In this context, 'digital sovereignty' refers to Europe's ability to act independently in the digital world and should be understood in terms of both protective mechanisms and offensive tools to foster digital innovation (including in cooperation with non-EU companies). See on that concept applied to the EU strategy, the report of the EU Political Strategy Center(EPCS), Rethinking Strategic Autonomy in the Digital Age, July 2019, available at https:// op.europa.eu/en/publication-detail/-/publication/889dd7b7-0cde-11ea-8c1f-01aa75ed71a1/language-en/format-RDF and F.Dezeuse and P.Timmers, Strategic Autonomy in the Digital Age, April,28,2022 available at: https:// policylabs.frontiersin.org/content/strategic-autonomy-in-the-digital-world. On digital sovereignty, read, among others, the excellent contribution of A.T. NORODOM, "Etre ou ne pas être souverain, en droit, à l'ère numérique" in Enjeux internationaux des activités numériques, C. Castets-Renard et alii (under the direction of), Larcier, Brussels, 2020, p. 21 and following.

[24] One weakness, however, that is often complained about is the level of European investment. In this respect, the figures quoted by the JRC report (M. Craglia (ed.), *Artificial Intelligence - A European perspective*, Publications Office of the European Union, Brussels, 3 December 2018, https://doi.org/10.2760/11251): "... United States, investments by GAFAM (private sector) and public authorities, DARPA (US Department of Defence Research Directorate: 7.5 billion dollars in 2020); China, for a volume of more than 20 billion; Europe (2.5 billion euros for 2018-2020), following the joint declaration of the Member States in April 2018 on their cooperation in the field of artificial intelligence. See on Global total corporate artificial intelligence (AI) investment from 2015 to 2022 https:// www.statista.com/statistics/941137/ai-investment-and-funding-worldwide/ and the report Stanford Institute for Human-Centered Artificial Intelligence (HAI, 2023, AI Index report available at : https://aiindex.stanford.edu/report/.

[25] www.usinenouvelle.com/etats-unis.

[26] Following the figures published the 14th of January 2021 by Analytica Insights (http://www.analyticsinsight.net/artificial-intelligence-investment-by-top-10-countries/.), "*China has always held ambitions significantly high for becoming AI Superpower of the world. In light of this goal the State Council of People's Republic of China has declared to become a $150 billion AI global leader by 2030. The United States of America is giving a tough competition to China in terms of becoming AI superpower. With the well-established tech culture in US, the country has been benefited with $10 billion venture capital channeling in direction of AI. But the future of AI has become unclear and is expected to decline due to recent country activities including reduced funding for AI, acceleration in education costs and strict immigration restrictions for international research professionals".* In comparison*, "The government of France is investing $1.8 billion in AI researches until 2022. The French AI initiatives will zoom into data with strategy to make private companies publicly release their data for utilizing it as AI use-cases.* ».

[27] *White Paper on artificial intelligence* (op. cit, 4): "However, the amount of investment in research and innovation in Europe remains well below the public and private investment in this field in other regions of the world. Some €3.2 billion was invested in AI in Europe in 2016, compared to about €12.1 billion in North America and €6.5 billion in Asia".

countries' companies (extraterritorial effect of the EU regulation)[28]. Indeed, the AI Act will impose to the high-risk AI systems a procedure of conformity assessment by accredited bodies and, in case of success, the delivery of a certificate[29]. It is clear that these certification systems are a major challenge for the creation of a European market for products and services that comply with regulatory requirements and the promotion of European players on this market, with the hope that these certificates can also be an added value on export markets. It must be added that the system of voluntary codes of conduct or accreditations enacted by the GDPR is here abandoned at least for high-risk AI systems at the benefit of a mandatory system already existing in other legislations, notably as regards certain medical devices[30].

On that point we might speak about a certain EU legal 'imperialism', since indirectly the EU is imposing its rule towards private sector companies of third countries and thus introduces the obligation for these third countries to pay attention to the EU legislation as it was clearly asserted in the GDPR case. The previous case of GDPR about data protection might be taken as an example of the EU policy on that point[31]. The same provision imposing duties to all digital actors acting within Europe or targeting EU citizens are present in most of the regulations mentioned above. In the name of this enlargement of the scope of these regulations, by instance, this broadening of the scope *ratione loci* of the European texts reflects the European will to use the regulatory tool to guarantee the protection of persons residing in Europe and, consequently, their trust in the AI tool developed or used there. Beyond that, it is an attempt to export the European regulatory model, insofar as the penetration of the European space by companies located outside Europe obliges them to obey the rules that prevail there and invites them to avail themselves of the added value of these rules regarding all their markets. For instance, the EU Commission, the 17th of December 2023, has decided to make investigation against Tik Tok and overall Twitter for non-respect of the Digital Service Act[32]. As regards Twitter, the Commission's investigation would focus on alleged breaches under four articles of the DSA related: 1. to dissemination of illegal content,

---

[28] Article,2.1 provides: "This Regulation applies to: *(a)*providers placing on the market or putting into service AI systems in the Union, irrespective of whether those providers are established within the Union or in a third country; *(b)*users of AI systems located within the Union; *(c)*providers and users of AI systems that are located in a third country, where the output produced by the system is used in the Union;"

[29] Article 31 of the AI Act enunciates; "The application for notification shall be accompanied by a description of the conformity assessment activities, the conformity assessment module or modules and the artificial intelligence technologies for which the conformity assessment body claims to be competent, as well as by an accreditation certificate, where one exists, issued by a national accreditation body attesting that the conformity assessment body fulfils the requirements laid down in Article 33. Any valid document related to existing designations of the applicant notified body under any other Union harmonization legislation shall be added." Article 43 and ff. are describing the conditions and the duration of the certificates issued by the accredited bodies.

[30] Regulation (EU) 2017/745 of the European Parliament and of the Council of 5 April 2017 on medical devices, amending Directive 2001/83/EC, Regulation (EC) No 178/2002 and Regulation (EC) No 1223/2009 and repealing Council Directives 90/385/EEC and 93/42/EEC.

[31] The two Schrems cases judged by the EUCJ might be quoted in that sense by putting an "emphasis on the importance of examining the practices of third-country public authorities in the exporters' legal assessment to determine whether the legislation and/or practices of the third country impinge – in practice – on the effectiveness of the *Art. 46 GDPR transfer tool*," That requirement imposes to the EU personal data exporter and the foreign importer are checking if enforceable data subject rights and effective legal remedies for data subjects are available in the country data is sent to.

[32]About that EU Commission decision as regards Twitter, read notably: https://www.euronews.com/my-europe/2023/12/18/brussels-launches-legal-action-against-musks-x-over-illegal-content-disinformation.

including measures taken to identify and swiftly remove illegal content and to allow users to flag such content; 2. to the absence of measures taken to combat information manipulation, such as the so-called "community notes"; 3. to take measures in order to increase transparency, amid suspected failures to provide researchers with adequate access to X's publicly accessible data; 4. to have developed deceptive design of X's user interface, such as the so-called "blue checkmark," a paid-for continuation that designates a user as an 'active, notable, and authentic' member of the platform.

9.2. Understanding the AI Act

9.2.1. Two key words: Excellence and trust.

But let us come back to the AI act. The "Third way", according to the famous U. van der Leyen discourse already quoted, is based on the two terms used in the title of the White Paper on artificial intelligence: on the one hand, **Excellence**, which characterises the quality of AI applications, and of research that supports their design, and on the other hand, **Trust**, which is necessary for the social acceptability of innovative digital developments, regardless of their field: education, health, mobility, justice, public affairs, etc.

Two guiding principles might be underlined on that point. First, Trust means to give the priority to the Human vis à vis the machine: It is a question of putting human at the centre of the digital development. At the same time, trust needs a 'Responsible Research and Innovation' (RRI[33]), a solid legal framework for accountable operators that allows for responsible innovation, without stifling innovation. This need for reconciliation is not necessarily contradictory. From one part, there is what we might call the precautionary principle, demanding attention to the risks and definitively limiting these risks; and, from the other part, what certain consider as another principle: the principle of innovation[34], which consecrates the support from the public authorities of innovations and their benefits for society. This balance, however, is not obvious to establish.

This policy, which is particularly explicit about AI systems, seeks to reconcile respect for European ethical values without concealing the fact that this respect has an economic objective: i.e., the creation of a strong and sovereign European market, notably through the creation of European labels or certificates. As Ms Vestager, the EU Commissioner, emphasised when presenting the proposal for an "AI Act" Regulation, the aim of this text is to implement the very principles of excellence and trust:

> *"On Artificial Intelligence, trust is a must, not a nice to have. With these landmark rules, the EU is spearheading the development of new global*

---

[33] **Responsible Research and Innovation (RRI)** is a concept developed by the European Union's Framework Programmes. It targets scientific research and technological development processes that take into account effects and potential impacts on the environment and society. About this concept and its implications, read: European Commission (2013). "Options for Strengthening Responsible Research and Innovation - Report of the Expert Group on the State of Art in Europe on Responsible Research and Innovation" (PDF). Publications Office. doi:10.2777/46253. Retrieved 24 June 2014.

[34] "The Innovation Principle is a tool to help achieve EU policy objectives by ensuring that legislation is designed in a way that creates the best possible conditions for innovation to flourish. The principle means that in future when the Commission develops new initiatives it will take into account the effect on innovation. This will ensure that all new EU policy or regulations support innovation and that the regulatory framework in Europe is innovation-friendly." (About this nascent principle and its enactment in certain EU documents, see: "Ensuring EU legislation supports innovation" available at https://research-and-innovation.ec.europa.eu/law-and-regulations/ensuring-eu-legislation-supports-innovation_en).

> *norms to make sure AI can be trusted. By setting the standards, we can pave the way to ethical technology worldwide and ensure that the EU remains competitive along the way. Future-proof and innovation-friendly, our rules will intervene where strictly needed: when the safety and fundamental rights of EU citizens are at stake"*[35].

The purpose of this major document is, according to the Commissioner, fourfold: 1) Ensure that AI systems placed on the EU market and used are safe and respect existing fundamental rights legislation and EU values; 2) ensuring legal certainty to facilitate investment and innovation in AI; 3) strengthen the governance and effective implementation of existing legislation on fundamental rights and safety requirements for AI systems; 4) facilitate the development of a single market for legal, safe and trustworthy AI applications, and prevent market fragmentation.

9.2.2. The scope: from AI machine learning to 'foundation models'

Having that strategy in mind, let us analyse the main points of the AI Act. The initial proposal dated April 21, 2021[36] has been since deeply amended both by the EU Council of Ministers (December, 22) and more recently by the EU Parliament (June, 23). The final compromise[37] has been approved the 8th of December and must still be translated in a final text and the adoption of the Act, hopefully at the beginning of 2024[38]. A preliminary point intends to fix the scope of the regulation by giving a definition for AI. According to the first draft, three technologies were included: the expert system translating in programme the causal human reasoning; the statistical applications; and finally the machine learning system, supervised or not, deep or not.

When they re-examined the AI Act proposal tabled by the Commission on 21 April 2021, both the Council and Parliament took the opportunity, first to reduce the definition only to machine-learning systems according with the OECD AI definition[39], since the other

---

[35] European Commission, Europe fit for the Digital Age: Commission proposes new rules and actions for excellence and trust in Artificial Intelligence, 21 April 2021, available at: <https://ec.europa.eu/commission/presscorner/detail/en/IP_21_1682>.

[36] European Commission, Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence legislation and amending certain Union legislative acts COM(2021), Brussels, 21 April 2021, 206, final {SEC(2021) 167 final} - {SWD(2021) 84 final} - {SWD(2021) 85 final}.

[37] The Parliament and the Council reached a provisional agreement on the AI Act. 14 June 2023 – The European Parliament adopted its negotiating position on the AI Act, with 499 votes in favor, 28 against, and 93 abstentions. 6 December 2022 – The Council of the EU adopted its common position ('general approach') on the AI Act. Parliament and Council negotiators reached a provisional agreement on the Artificial Intelligence Act, finally, the 9th of December (see the announcement of the agreement by the EU Parliament: https://www.europarl.europa.eu/news/en/press-room/20231206IPR15699/artificial-intelligence-act-deal-on-comprehensive-rules-for-trustworthy-ai.)

[38] Since the date of writing, a final version of the EU AI Act has been adopted by the European Parliament. This final version could not be considered in time for the closing of this volume. However, the editors and author still consider the reflections contained herein relevant for analysis.

[39] "An AI system is a machine-based system that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments. Different AI systems vary in their levels of autonomy and adaptiveness after

technologies, especially the expert-systems, did not raise the same concerns due to their transparent functioning and their intrinsic limitations; and second, to address the specific issues raised by the 'General Purpose AI models' (GPAI) or foundation models, as designated under the EU Parliament compromise[40]. As regards these "foundation models" like "transformers" or LLM (Language large models: Ernie, BERT, GPT), the question was the inadequacy of the definition given in the initial Commission's text to cover these so-called "general purpose" AI systems. The initial text defines AI systems in terms of the specific purposes pursued, in line with the GDPR requirement. However, as already stated, 'foundation models' pursue a general purpose which, admittedly, will allow the emergence of multiple applications linked to specific purposes, such as offering companion chatbot services or designing publicity campaigns.

Both the Council and the Parliament are therefore amending the definition of artificial intelligence and are also proposing a definition of the concept of 'foundation model'. Using the wording used by the European Parliament in the compromise agreement published on 14 June 2023, article 3.1 defined artificial intelligence by deleting the reference to "*for a given set of human-defined objectives" as follows: "'artificial intelligence system' (AI system) means a machine-based system that is designed to operate with varying levels of autonomy and that can, for explicit or implicit objectives, generate outputs such as predictions, recommendations, or decisions, that influence physical or virtual environments*;". This extension explicitly encompasses GPAIs. These generative AI models might be defined as follows: "general purpose AI model' *means "an AI model, including when trained with a large amount of data using self-supervision at scale, that displays significant generality and is capable to competently perform a wide range of distinct tasks regardless of the way the model is placed on the market and that can be integrated into a variety of downstream systems or applications*."[41]

deployment." (OECD, Recommendation of the Council on Artificial Intelligence, OECD/LEGAL/0449, adopted by the Council of Ministers, May 22, 2019 (https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449.).

[40] From the outset, we would like to introduce a distinction between "generative AI" models and applications to avoid the confusions too often heard and to explain the proposals currently being discussed in the framework of the proposal for a European regulation on artificial intelligence, known as the AI Act1. Our discussion only analyzes the regulation of models, i.e. "generic" generative systems, "foundation models", which make possible a wide range of applications that we won't go into in this brief. To come back to models, we are talking about "transformers models" which allow an AI system to exploit relationships between elements of a data set without these relationships being dictated beforehand by humans, "Large Language models" (LLMs) about machine-operated aggregations of textual elements, "multimodal languages", capable of working on texts at the same time, images and sounds. Without being exhaustive, far from it, Google has developed BERT ("Bi-directional Encoder Representations from Transformers"); ChatGpt (Generative Pre-Trained Transformer) was developed by Open AI and is backed by Microsoft; Baidu, a Chinese Bigtech company, is developing ERNIE ("Enhanced Representation through kNowledge Integration"); Meta, "Open Pre-Trained Transformer" (OPT-175B) and LLamA. About these LLM, read the excellent study published by the OECD:, « AI Language Models Technological, Socio-Economic and Policy Considerations », OECD Digital Economy Papers, avril 2023, n° 352 and the interesting UNESCO report: « Foundation Models such as ChatGPT through the prism of the UNESCO Recommendation on the Ethics of Artificial Intelligence, UNESCO, June 2023 ») analyzing how the UNESCO recommendation on AI ethics might be applied to "foundations models" *in casu* Chat GpT

[41] According with the COREPER definition (art. 3.44 b)) published the 2nd of February 2024.

9.2.3. The risk-based approach

What about the content of the AI regulation? The first point is definitively the **risk approach** adopted by the text after other texts. The main idea is according with the proportionality principle to distinguish the provisions according with the gravity of the risks created by the product, the service, or the application[42]. The same idea runs through the "AI ACT". The text distinguishes five categories of AI systems, to which correspond different types of obligations. First, the proposal sets out, first, the prohibition of illegal practices of artificial intelligence[43] corresponding to an intolerable level of risk (Art. 5), like social ranking, people manipulation, facial recognition in public areas except for police activities, emotion recognition (but only in the workplace and education sector)[44]. Second, the text establishes a system of control and management of high-risk AI systems, which are listed in two annexes that may be amended by the Commission, one referring to existing texts already enacted by the EU (for instance, regarding the medical, aeronautical or financial sectors), and the other one, defining different purposes considered as presenting 'high risks'. As regards this second list, I pinpoint the AI applications for selecting employees or for evaluating students but also applications by the law enforcement agencies or used for influencing the outcomes of elections. The final compromise grants rights to citizens to launch complaints against AI systems and to receive explanations about decisions generated by AI high-risks systems (more or less a copy of the GDPR article 22).

The so-called 'foundation models' category receives a specific regulation inspired by that of 'high risk systems', but a bit lighter. Transparency obligations about the training data will apply to all these models and 'the use of AI generative systems will have to be recognisable by the users. Furthermore, the text fixes additional obligations for 'high impact' foundation models. Therefore, a new annex will provide criteria (to be adapted to the technological evolution): notably, the large amount of data used for training the system, the number of business users and the complexity of the model's parameters[45]. A fourth category imposes specific obligations for lack of transparency of certain hidden applications "*in particular when ultra-realistic dialogue or video tricks are used*" (the text targets in particular the

---

[42] The genuine risk-based approach leads to the creation of new obligations when certain criteria proposed by the regulation indicate that higher risks are present. This approach is already used, but in a very limited way, in the provisions of the RGPD: Article 35 reserves the obligation to carry out an impact assessment only to processing operations presenting a "high risk" to the rights and freedoms of natural persons. The notion of "high risk" remains unclear under GDPR, even if the EDPB has tried to clarify it. The Regulation on medical devices similarly distinguishes between different classes of products and services according to the purpose of their use and the risks related to health and safety, and subjects "high risk" classes of products to conformity assessment procedures. See also the DSA which distinguishes different categories of intermediaries', since simple hosting services till 'very large platforms', according to the gravity of the risks linked with their interventions.

[43] For example, subliminal message manipulation systems, the exploitation of vulnerabilities (manipulation by deepfakes?), the use by the public sector of "social ranking" systems leading to potential discrimination between individuals or groups, biometric systems operating in real time and remotely, placed in public places (e.g. facial recognition systems).

[44] It seems that on that point, the long list of prohibited practices proposed by the EU Parliament was not accepted by the Council and that at the end, the text will be close to the Council's position. By instance, the new text will open the possibilities for law enforcement agencies to use remote biometric identification systems but these systems will be considered as high-risk systems.

[45] Non systemic pretrained models might avoid all these obligations "if they are made accessible to the public under a license that allows for the access, usage, modification, and distribution of the model, and whose parameters are made publicly accessible".

deepfakes); and, finally, a fifth category, for other applications presenting a low-level risk, left to self-regulation by the market.

The second point to consider is linked with the first: it deals with the enlargement of the risks to be considered. The "AI Act", or rather the work of the EU *High-Level Group of Experts on AI*,[46] to which the proposal constantly refers, broadens the risks to be taken into consideration when assessing AI applications. Thus, in addition to the risks to our individual freedoms (data protection, dignity or freedom of expression), there is the need to take into consideration the so-called collective risks specific to a group of people identified *a priori* or not, or the risks of undermining social justice by discriminating certain categories of people and, beyond that, the societal risks, such as the potential threats to the environment, competition, democracy, and respect for the rule of law.

Another consequence of the risk approach is the reference to the need to have a preventive assessment by the main actors of the AI supply chain, the providers and the deployers, before putting their models or applications or their uses of applications on the market. By deployers, the text means the AI system users, which are using AI applications in the context of their business or administrative activities (by instance a bank using an AI system for evaluating the creditworthiness of their customers or fiscal administrations, for detecting fiscal fraud). The risk-based approach fully justifies the shift from a classic legal drafting – based on the definition of behavioural content to be respected and, in the event of non-compliance, on the repression or *a posteriori* sanctions of breaches of the regulations – to an *a priori* approach based on the obligation for the companies or administrations to assess the risks linked to their development or use of technological products or services, i.e. to set up a risk assessment procedure and monitor compliance with this procedure.

The obligation to conduct a prior assessment has been already imposed in the context of the GDPR but was only relative to data protection issues and poorly regulated as regards the procedure to be followed. The proposed "AI Act" enlarges the scope of the risks to be assessed as developed above[47] but also develops this procedure at leisure, defining its stages, its content, insisting on the participation of all the interested parties, etc. This approach is to be commended, although it is administratively more cumbersome and can only be justified in cases of high-risk systems and of foundation models. It imposes the obligation for the companies to set up a quality management procedure[48] and body, to include in the process the

---

[46] High-Level Expert Group on AI (HLGE), *Ethical guidelines for trustworthy AI*, 8 April 2019, No. 67, text available at: Ethics guidelines for trustworthy AI - Publications Office of the EU (europa.eu).

[47] See, our reflections in Y. POULLET, The Data Protection Impact Assessment or rather the Privacy Impact Assessment, a Revolution with a Future in the Age of Artificial Intelligence? in *Artificial intelligence Law*, Céline CASTERS-RENARD and J. EYNARD (eds), Bruylant, 2023, p.627-647: " *Precisely, the proposal concerning AI systems clearly broadens the debate. It intends to consider not only the risks incurred by our individual liberties or jeopardizing our interests as consumers (see the systems of recommendation of goods or products) but also the risks of discrimination and non-respect of the values of social justice or even the societal risks, such as environmental issues, the attacks on the rule of law and democracy.*" (p.637)

[48] The article 9 defines and enunciates the purposes of the quality management: "1. A risk management system shall be established, implemented, documented and maintained in relation to high-risk AI systems. *2.* The risk management system shall consist of a continuous iterative process run throughout the entire lifecycle of a high-risk AI system, requiring regular systematic updating. It shall comprise the following steps: *(a)*identification and analysis of the known and foreseeable risks associated with each high-risk AI system; *(b)*estimation and evaluation of the risks that may emerge when the high-risk AI system is used in accordance with its intended purpose and under conditions of reasonably foreseeable misuse; *(c)*evaluation of other possibly arising risks

main stakeholders[49] and to identify the risks and to mitigate them, including the obligation to pinpoint the residual risk and to provide a continuous assessment of the AI system. The control by accredited bodies (the notifying bodies), the delivery of certificates by them and the publication of the high-risk system in a publicly accessible register are required.

This assessment aims to ensure the respect of the ethical values asserted by the AI Act itself (article 4): "Operators falling under this Regulation shall make their best efforts to develop and use AI systems or foundation models in accordance with the following general principles establishing a high-level framework that promotes a coherent human-centric European approach to ethical and trustworthy Artificial Intelligence, which is fully in line with the Charter as well as the values on which the Union is founded". Six values are to be considered: 1. 'human agency and oversight' including the respect of human dignity and personal autonomy; 2 'technical robustness and safety'; 3. 'privacy and data governance'; 4. 'transparency'; 5 'diversity, non-discrimination and fairness'; 6.'social and environmental well-being'.

9.2.4. A strong governance model and sanctions.

For respecting theses values, for instance, the deployer of high-risk AI systems (article 29) shall ***take appropriate technical and organisational measures to ensure they*** use such systems in accordance with the instructions of use accompanying the systems and with the legislative requirements. This means: 1. implement effective human oversight by qualified and competent human beings; 2. ensure that data used are relevant and sufficiently representative; 3. without undue delay, inform and cooperate with the provider and the national supervisory authority when they identify the presence of a risk and even to suspend the use of the system in case of risk occurrence 4. keep the logs automatically generated by the use of the system; 5. if required by GDPR[50], carry out a data protection assessment and publish a summary; 6 in case of a system used for preparing or making decisions, inform the data subject about the use of the system and the right to a human explanation.

The recourse to existing standards as regards notably the training activities or the acceptable parameters or even the methodology to be followed during the assessment procedure will facilitate the task of the evaluators of the compliance of the AI systems with the AI Act requirements. In order to encourage the recourse to EU standards, the reference to standards is considered by the EU proposal as a way to ensure a presumption of compliance of the AI

based on the analysis of data gathered from the post-market monitoring system referred to in Article 61; *(d)*adoption of suitable risk management measures in accordance with the provisions of the following paragraphs."

[49] The need to have an evaluation of the AI system, open to the participation of the different stakeholders is asserted at different levels. An amendment introduced by the EU Parliament and likely taken again by the final proposal underlines that high-risk AI systems used in the context of relationships with workers must be discussed with the representative of these latter. More broadly, the text of the EU Parliament provides: "*In the course of the impact assessment, the deployer, with the exception of SMEs, shall notify national supervisory authority and relevant stakeholders and shall, to best extent possible, involve representatives of the persons or groups of persons that are likely to be affected by the high-risk AI system, as identified in paragraph 1, including but not limited to: equality bodies, consumer protection agencies, social partners and data protection agencies, with a view to receiving input into the impact assessment."*

[50] The final compromise between EU Parliament and the EU Council maintains the existence of the GDPR procedure as a mandatory element distinct from the assessment of other risks.

applications with the legal obligations[51]. Always as regards the standards, the AI Act proposal of the EU Parliament also introduced a revision of the composition and of the procedures of the EU standardisation bodies to take into account the new issues raised by this technology. According to that proposal, the final compromise of the AI act is pushing forward the EU standardisation[52]. It might be added that in that context, the EU is claiming a more open standardisation procedure, which includes the possible intervention of the different stakeholders[53]. That requirement to enlarge the composition or at least the procedure of the standardisation bodies (CEN, CENELEC, ETSI) is to be underlined since it will deeply modify the culture of the EU standardisation bodies, traditionally joining together only engineers and focusing only on technical questions. In conclusion, if most of the standards available in Europe are international ones adopted within ISO, in the future, the EU Parliament claim will lead to more independent EU standards and furthermore collectively discussed.

Fifth point, in order to promote innovation, the Members States might exonerate the AI actors from certain obligations in the context of so-called 'sandbox' legislations: *"AI regulatory sandboxes shall, in accordance with criteria set out in Article 53a, provide for a controlled environment that fosters innovation and facilitates the development, testing and validation of innovative AI systems for a limited time before their placement on the market or putting into service pursuant to a specific plan agreed between the prospective providers and the establishing authority".* Certain guarantees will surround these initiatives[54]. The final compromise confirms the possibility to develop and train innovative AI systems in real-world conditions before their placement on the market. Specific provisions will be proposed to alleviate the administrative burdens for smaller companies.

Furthermore, it must be emphasised that the AI Act multiplies the bodies in charge to make the legislation effective and develop a new governance architecture. Beyond that

---

[51] See article 40.1 modified partly by the Amendment 438 of the EU Parliament text: "High-risk AI systems and foundation models which are in conformity with harmonised standards or parts thereof the references of which have been published in the Official Journal of the European Union in accordance with Regulation (EU) 1025/2012 shall be presumed to be in conformity with the requirements set out in Chapter 2 of this Title or Article 28b, to the extent those standards cover those requirements."

[52] According to the 'New approach to enable global leadership of EU standards promoting values and a resilient, green and digital Single Market' adopted by the EU Commission, February, the 2nd of 2022. On that EU strategy on standardization, see: https://single-market-economy.ec.europa.eu/news/new-approach-enable-global-leadership-eu-standards-promoting-values-and-resilient-green-and-digital-2022-02-02_en. Thierry Breton said: **"Technical standards are of strategic importance. Europe's technological sovereignty, ability to reduce dependencies and protection of EU values will rely on our ability to be a global standard-setter. With today's Strategy, we are crystal-clear on our standardisation priorities and create the conditions for European standards to become global benchmarks. We take action to preserve the integrity of the European standardisation process, putting European SMEs and the European interest at the centre".**

[53] Under the proposed article 40.1.c) of the EU Parliament, it was provided that: "The actors involved in the standardisation process shall take into account the general principles for trustworthy AI set out in Article 4(a), seek to promote investment and innovation in AI as well as competitiveness and growth of the Union market, and contribute to strengthening global cooperation on standardisation and taking into account existing international standards in the field of AI that are consistent with Union values, fundamental rights and interests, and **ensure a balanced representation of interests and effective participation of all relevant stakeholders** in accordance with Articles 5, 6, and 7 of Regulation (EU) No 1025/2012."

[54] "Establishing authorities shall provide guidance and supervision within the sandbox with a view to identify risks, in particular to fundamental rights, democracy and rule of law, health and safety and the environment, test and demonstrate mitigation measures for identified risks, and their effectiveness and ensure compliance with the requirements of this Regulation, …" (article 53.1)

consideration, the text refers to a lot of other bodies in charge of certain competences. Therefore, we have already spoken about the notifying bodies, accredited private or public companies in charge of the audit and control of the conformity of the high-risk systems. The AI Act asserts the principle of a free choice of the notifying body by the provider or the deployer. A second body is set up by the regulation; the national supervising authority[55] which oversees the control of the respect of the legislation. On that point, we underline the need for a cross-cutting approach since the risks linked with AI development, as previously said, are not only a question of data protection and individual freedoms but covers other issues like discrimination, competition, environment, and democracy. The proliferation within EU of administrative authorities in charge of these different aspects raises difficulties when it comes to analysing the impact of a technology like AI in a cross-cutting manner. In our opinion, the difficulty might be met only by institutionalising the creation of forums for dialogue between these different bodies, without which there is a risk of contradictory interventions or even rivalry between authorities.

In order to further increase the effectiveness of the regulatory texts and to ensure their rapid adaptation to the needs of technological development, the text confers large powers on the Commission, either to monitor the application of the regulations in the form of reports in particular, or to adopt delegated acts pursuant to the text of the Regulation, for instance by reviewing the scope of the AI Regulation, by completing the list of high-risk systems, by defining the "high impact" foundation models, etc. The Commission itself is assisted by an advisory and consultative body: the AI Office with competences such as overseeing the most advanced standards and suggesting new rules or actions to the Commission, like fostering standards and testing practices. A scientific panel of independent experts will advise this AI office as regards the methodologies to be used for evaluating the capabilities and the risks of 'foundation models' especially as regards the designation of "high impact foundation models" and the security measures to be taken. It seems that under the final compromise, this AI Office will be composed by representatives of the different stakeholders, as claimed by the EU Parliament[56]. An AI Board is set up with representatives of the different member states. Its role is to act as a coordination platform and as an advisory board to the Commission. It must be added that an advisory forum with representatives of the different stakeholders (start-ups, industry representatives, civil society, and academia) will provide technical expertise to this Board.

The final compromise adopts provisions about penalties in case of violations of the AI Act. A percentage of the offending company's global annual turnover in the previous financial year or a predetermined amount are foreseen depending on the gravity of the violation, except in case of infringements by SME or start-ups. Finally, the right for natural and overall legal persons to address a complaint to the EU market surveillance authority is enacted.

---

[55] The choice of the supervising authority is not obvious since a certain number of 'independent' administrative authorities might revendicate this role: DPA (Data Protection Authority), Commission for equal opportunities, Competition authority, Media and Audiovisual Commission, etc. In the Netherlands, the DPA has been provisionally chosen for this task. The UK model creating an umbrella organization joining together all the independent administrative authorities, dealing with AI systems issues seems quite interesting.

[56] The Amendment 122 of the EU Parliament text dated from June the 14th introduced a new Recital 76 that clearly stipulated: "…. Stakeholders should formally participate in the work of the AI Office through an advisory forum that should ensure varied and balanced stakeholder representation and should advise the AI Office on matters pertaining to. In case the establishment of the AI Office prove not to be sufficient to ensure a fully consistent application of this Regulation at Union level as well as efficient cross-border enforcement.". At this moment, we are not aware that this amendment has been rejected by the final compromise.

9.3. Conclusions

The following points might be laid down from the previous reflections:

- AI Act must be considered as an element of a global, sovereign EU data strategy, and a coherent legal framework through different regulations aiming at ensuring the EU 'third way' based on "trust", and "excellence".
- With the AI Act, the EU might be considered as a leader in the field of rulemaking and not necessarily in terms of business and market AI development and commercialization. By this legislative approach, the EU institutions hope, following the GDPR previous model, to impose swiftly a global legislative standard as regards AI systems. Another expected benefit of this approach is to avoid a fragmented EU market.
- The EU text is balancing between a **precautionary** approach and a **competitive** one pushing forward innovation. The first one is illustrated by the multiple obligations imposed to the AI systems providers and deployers and the existence of multiple bodies in charge supervising the AI Act. The second one explains both the risk approach exempting a large number of AI systems from all the administrative burdens and the potential existence of the "sandbox legislations".
- The EU regulatory approach is based mainly on the accountability principle. It is the role of the AI developers and users to analyse themselves the risks linked to their systems and if needed, to mitigate the risks. At the same time, the EU regulations are set up as a strong governance architecture by reinforcing the enforcement powers of the EU Commission. It must be added that at the two levels (micro and macro), the EU insists on the importance of developing an inclusive participation of all interested stakeholders, including the civil society, the DPA, the trade unions, the consumers' associations, etc.
- The existence and the role of specific EU "standards" in the development of AI systems and the lack of standards as regards especially generative models and applications must be pinpointed. On that point, it might be difficult to imagine that global technical and managerial standards would not be defined by international public (ITU) private (ICANN, Web 3C, IETF...) or mixed public-private (ISO) organisations.
- Undoubtedly, there will be an urgent need for further reflections about legal aspects, such as environmental protection, liability, intellectual property, privacy protection, competition, discrimination and freedom of expression linked with the development of foundation models and their applications. The question of liability for the damages caused by high-risk systems is presently discussed by the EU authorities in the context of distinct legislative initiatives[57].

We do hope that the EU approach of promoting a human-centric approach and asserting the need to respect ethical values as regards AI development and applications might be a good example for our digital world by offering strong safeguards to our liberties and democracies against potential public or private abuses of technology. However, at the same time, it is not obvious that such EU model, which may be excessively administrative, will be accepted by private actors and will be the right way to achieve international consensus.

[57] See particularly, the proposal for a Directive (EU) of the European Parliament and of the Council on adapting non-contractual civil liability rules to artificial intelligence (AI Liability Directive), **OJ L 409, 4.12.2020, p. 1–27.**

References

ANALYTICA INSIGHTS. Artificial Intelligence Investment by Top 10 Countries, 14 jan. 2021. Available at: <http://www.analyticsinsight.net/artificial-intelligence-investment-by-top-10-countries/>.

Brussels launches legal action against Musk's X over illegal content disinformation. Euronews, 18 dec. 2023. Available at: <https://www.euronews.com/my-europe/2023/12/18/brussels-launches-legal-action-against-musks-x-over-illegal-content-disinformation>.

COMMUNICATION FROM THE COMMISSION TO THE EUROPEAN PARLIAMENT, THE COUNCIL, THE EUROPEAN ECONOMIC AND SOCIAL COMMITTEE AND THE COMMITTEE OF THE REGIONS. A European strategy for data, COM/2020/66, Brussels, 19 feb. 2020.

CRAGLIA, M. (Ed.). Artificial Intelligence - A European perspective. Publications Office of the European Union, Brussels, 3 dec. 2018. Available at: <https://doi.org/10.2760/11251>.

DEZEUSE, F.; TIMMERS, P. Strategic Autonomy in the Digital Age, 28 apr. 2022. Available at: <https://policylabs.frontiersin.org/content/strategic-autonomy-in-the-digital-world>.

Ensuring EU legislation supports innovation. Available at: <https://research-and-innovation.ec.europa.eu/law-and-regulations/ensuring-eu-legislation-supports-innovation_en>.

EU COMMISSION. New approach to enable global leadership of EU standards promoting values and a resilient, green and digital Single Market, 2 feb. 2022.

EU POLITICAL STRATEGY CENTER (EPCS). Rethinking Strategic Autonomy in the Digital Age, July 2019. Available at: <https://op.europa.eu/en/publication-detail/-/publication/889dd7b7-0cde-11ea-8c1f-01aa75ed71a1/language-en/format-RDF>.

EUROPEAN COMMISSION. Europe fit for the Digital Age: Commission proposes new rules and actions for excellence and trust in Artificial Intelligence, 21 apr. 2021. Available at: <https://ec.europa.eu/commission/presscorner/detail/en/IP_21_1682>.

EUROPEAN COMMISSION. Options for Strengthening Responsible Research and Innovation - Report of the Expert Group on the State of Art in Europe on Responsible Research and Innovation. Publications Office, 2013. doi:10.2777/46253.

EUROPEAN COMMISSION. Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence legislation and amending certain Union legislative acts COM(2021), Brussels, 21 apr. 2021, 206, final {SEC(2021) 167 final} - {SWD(2021) 84 final} - {SWD(2021) 85 final}.

EUROPEAN COMMISSION. USA-China-EU plans for AI: where do we stand? Jan. 2018. Available at: <https://monitor-industrial-ecosystems.ec.europa.eu/reports/other-reports/usa-china-eu-plans-ai-where-do-we-stand>.

EUROPEAN PARLIAMENT. Resolution of 9 September 2010 on Better Lawmaking 14 (P7TA(2010)0311.

EUROPEAN SOCIAL AND ECONOMIC COMMITTEE (EESC). Self-and Co-regulation. Available at: <https://www.eesc.europa.eu/sites/default/files/resources/docs/auto_coregulation_en--2.pdf>.

HIGH-LEVEL EXPERT GROUP ON AI (HLGE). Ethical guidelines for trustworthy AI, 8 apr. 2019, No. 67. Available at: <https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai>.

Le RGPD face aux défis de l'intelligence artificielle, Coll. Cahiers du CRIDS n° 48, Larcier, Bruxelles, 2020, p. 15 e seg.

New approach to enable global leadership of EU standards promoting values and a resilient, green and digital Single Market. Available at: <https://single-market-economy.ec.europa.eu/news/new-approach-enable-global-leadership-eu-standards-promoting-values-and-resilient-green-and-digital-2022-02-02_en>.

NORODOM, A. T. "Etre ou ne pas être souverain, en droit, à l'ère numérique", in: CASTETS-RENARD, C. et al. (Dir.). Enjeux internationaux des activités numériques. Brussels: Larcier, 2020, p. 21.

OECD. AI Language Models Technological, Socio-Economic and Policy Considerations. OECD Digital Economy Papers, n° 352, April 2023.

OECD. Recommendation of the Council on Artificial Intelligence, OECD/LEGAL/0449, adopted by the Council of Ministers, 22 may 2019. Available at: <https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449>.

POULLET, Y. Governance Challenges: First Lessons from the WSIS - An Ethical and Social Perspective. In: Innovation, Legitimacy, Ethics and Democracy. Liber Amicorum J. Berleur, PUN, Namur, 2007.

POULLET, Y. How to regulate the Internet: New paradigms for Internet Governance? In: WALDEN, I.; HÖRNLE, J. (Eds.). E-Commerce Law and practice in Europe. Cambridge: Woodehead, 2007. Section 1, Chap. 2.

POULLET, Y. The Data Protection Impact Assessment or rather the Privacy Impact Assessment, a Revolution with a Future in the Age of Artificial Intelligence? In: CASTERS-RENARD, C.; EYNARD, J. (Eds.). Artificial intelligence Law. Bruylant, 2023, p. 627-647.

Proposal for a Directive (EU) of the European Parliament and of the Council on adapting non-contractual civil liability rules to artificial intelligence (AI Liability Directive), OJ L 409, 4 dec. 2020, p. 1–27.

Proposal for a regulation laying down harmonised rules on artificial intelligence, 21 apr. 2021, COM (2021)206 Final 2021/0106.

Proposal for a REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL establishing a common framework for media services in the internal market (European Media Freedom Act) and amending Directive 2010/13/EU, Brussels, 16 sep. 2022, COM(2022) 457 final.

Regulation (EU) 2017/745 of the European Parliament and of the Council of 5 April 2017 on medical devices, amending Directive 2001/83/EC, Regulation (EC) No 178/2002 and Regulation (EC) No 1223/2009 and repealing Council Directives 90/385/EEC and 93/42/EEC.

Regulation (EU) 2019/881 of the European Parliament and of the Council of 17 April 2019 on ENISA (the European Union Agency for Cybersecurity) and on information and communications technology cybersecurity certification and repealing Regulation (EU) No 526/2013 (Cybersecurity Act), PE/86/2018/REV/1, OJ L 151, 7 jun. 2019, p. 15–69.

Regulation (EU) 2022/1925 of the European Parliament and of the Council of 14 September 2022 on contestable and fair markets in the digital sector and amending Directives (EU) 2019/1937 and (EU) 2020/1828 (Digital Markets Act), PE/17/2022/REV/1, OJ L 265, 12 oct. 2022, p. 1–66.

Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market For Digital Services and amending Directive 2000/31/EC (Digital Services Act), PE/30/2022/REV/1, OJ L 277, 27 oct. 2022, p. 1–102.

Regulation (EU) 2022/868 of the European Parliament and of the Council of 30 May 2022 on European data governance and amending Regulation (EU) 2018/1724 (Data Governance Act), PE/85/2021/REV/1, OJ L 152, 3 jun. 2022, p. 1–44.

REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL on harmonised rules on fair access to and use of data (Data Act), Brussels, 15 nov. 2023, 2022/0047 (COD) PE-CONS 49/23.

STANFORD INSTITUTE FOR HUMAN-CENTERED ARTIFICIAL INTELLIGENCE. AI Index report, 2023. Available at: <https://aiindex.stanford.edu/report/>.

STATISTA. Global total corporate artificial intelligence (AI) investment from 2015 to 2022. Available at: <https://www.statista.com/statistics/941137/ai-investment-and-funding-worldwide/>.

Towards a New EU Regulatory Approach of the Digital Society, ERDAL, 2022, Vol. 1, p. 113 e seg.

UNESCO. Foundation Models such as ChatGPT through the prism of the UNESCO Recommendation on the Ethics of Artificial Intelligence, June 2023.

WHITE HOUSE OFFICE OF SCIENCE AND TECHNOLOGY POLICY. Blueprint for an AI Bill of Rights: Making Automated Systems Work for the American People, October 2022.

White Paper on artificial intelligence.

10. 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

This Chapter explore 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 investigates 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 analysed 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.

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, analysing, and interpreting data, to make decisions and take actions with a certain degree of autonomy (Russell, S., and Norvig, P., 2003). 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 behaviour, by allowing or denying some course of action, influencing the way we make decisions, or playing a crucial role in the way relevant decision for individuals and communities are made (Palladino, N., 2023).

If not managed properly, AI systems may be subject to a series of misuses 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 (Renda, A., 2019).

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[1], or the UNESCO Recommendation on the Ethics of Artificial Intelligence[2], 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.

---

[1] 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.

[2] https://unesdoc.unesco.org/ark:/48223/pf0000381137

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) (Floridi, L., Cowls, J., Beltrametti, M., et al., 2018). 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 between mandatory requirements and self-governance practices to safeguard people without hindering innovation (Turner, J., 2019).

Although a "by design" approach is deemed crucial to ensure the effective safeguarding of human rights and ethical concerns in the digital realm (Suzor, N., Dragiewicz, M., Harris, B., Gillett, R., Burgess, J., and Van Geelen, T., 2019), 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 (Shilton, K., 2015).

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" (European Commission, 2019) 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" (European Commission, 2019).*

The European Union har recently approved the 'Regulation Laying Down Harmonised Rules on Artificial Intelligence ', better known as 'Artificial Intelligence Act' (AIA), establishing regulatory requirements for AI systems.

This chapter 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 chapter 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.

10.1. 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" (Palladino, N., & Santaniello, M., 2021).

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 behaviour, 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 behaviour "to people or groups outside the acting entity who are nevertheless affected by it" (Risse, T., 2006) 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 the system and the subsequent effects of that conduct," (Wieringa, M., 2020) 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 involves different kind of actors at various stages (Schiff, D., Rakova, B., Ayesh, A., Fanti, A., & Lennon, M., 2020). So, we can think about internal accountability like a chain of duty and responsibilities between different internal stakeholders. Instead, external accountability requires the evaluation from an external forum such as public opinion or an authority such as a regulatory or certification body. External accountability is particularly important in the context of artificial intelligence.

Since artificial intelligence systems incorporate within their architecture social norms and assumptions about the nature of the world, external accountability helps to draw out embodied values and requires decision makers to justify their choices and the algorithmic systems' outputs in front of public reason, that means according to epistemic and normative standards which are acceptable to all reasonable people (Binns, R., 2018).

Most of the international frameworks on trustworthy AI focus on a narrow set of requirements to achieve accountability. These requirements serve both internal and external accountability purposes. The two dimensions 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:

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."[3]

[3] China Academy of Information and Communications Technology. (2019, June 1). *《人工智能行业自律公约(征求意见稿)》公开征求意见 [The Self-Regulatory Convention for the Artificial Intelligence Industry (Draft for Comments) is open for public comment]*. SECRSS. https://www.secrss.com/article/11099.

ii) Verification and Validation: AI systems' providers must furnish proof that their application operates accurately in line with anticipated performance standards. According to IEEE[4], "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."[5]

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 (Association for Computing Machinery US Public Policy Council, 2017), 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 (Amnesty International and Access Now, 2018).

The Artificial Intelligence Act is the first framework attempting to institutionalize the aforementioned dimensions of accountability within a binding piece of legislation. Subsequent paragraphs will undertake a more comprehensive exploration of the mechanisms drafted in the proposal to ensure accountability.

10.2. 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). The proposal has been amended by the Council (December 2022) and the European Parliament (June 2023), and after a final phase of the legislative process, called "trilogue", in which the three institutions negotiate to produce an agreed version of the text, it has been approved in May 2024 and published in the Official Journal of the European Union on July 12, 2024

This paragraph will first provide a brief overview of the approach and the main characteristics of the Artificial Intelligence Act. Then, we will explore the accountability mechanisms that have been foreseen in the Act.

In its 2020 *White Paper on Artificial Intelligence (European commission., 2020)*, the European Commission outlined the need to develop an "ecosystem of trust" to foster the

---

[4] IEEE. *Ethically Aligned Design First Edition*. 2019. Available At: <Https://Standards.Ieee.Org/Industry-Connections/Ec/Ead1e-Infographic/>.

[5] Université de Montreal. (2017, November 3). *Montreal Declaration for a Responsible Development of Artificial Intelligence*. La Recherche - Université de Montréal. https://recherche.umontreal.ca/english/strategic-initiatives/montreal-declaration-for-a-responsible-ai/

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 (Tambiama M., 2022) with the publication of the Communication on *Fostering a European Approach to Artificial Intelligence (European commission., 2020)*.

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 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[6]. 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 analyse 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 Act 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 AIA 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.

Moreover, the Act 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 AIA 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.72). 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

[6] A more comprehensive description of these figures is available in Article 3 of the Act.

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 (Art. 40).

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* advisory fórum (Art.41).

AI applications not included in the previous categories are considered limited or minimal risk. They are subject to some mandatory requirements (Art. 50) 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 (Art.56).

The AIA 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. 49, 71). 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 (Art.17, 1m).

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 (Art.10, Art.15).

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 (Art.9).

Several provisions in the AIA 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

(Art.11). 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 (Art.12).

However, the Act 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 AIA: the Conformity Assessment (see Chapter 5 of the Act). 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.74), 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 actions 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 actions 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[7].

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.77).

[7] These are the powers assigned to market surveillance authorities by Regulation 2019/1020.

Finally, the AIA foresees economic penalties for non-compliance with the prohibitions and requirements set in the regulation.

10.3. 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.

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 of the presence and adequacy of measures to ensure their safety and respect for fundamental rights by third-party bodies, including the vast majority of applications classified as high-risk under the same regulation.

Even if AI providers are sanctioned for causing harm due to their negligence in complying with AIA requirements, such measures represent an ex-post intervention. This approach undermines the ability to prevent harms proactively and may hinder the development of a ecosystem of trust around the AI technology, which is one of the primary objectives of the AIA.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, 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 (Palladino, N., & Santaniello, M., 2021) and attempt to soften the burdens on businesses.

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 behaviour(Palladino 2023).

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.

In this regard the Commission may ensure that expert groups supporting its decisions on standards and technical specifications implementing the AIA include more representatives from civil society, consumer organizations, trade unions, and non-governmental organizations alongside business, technical and academic experts. Moreover, it should prohibit experts with direct ties to AI providers from taking part in these evaluation processes where their affiliations might benefit from a favorable outcome. Enhanced and more transparent disclosure procedures may help identify potential conflicts of interest, fostering accountability and trust in the evaluation process.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. Unfortunately the establishment of the AI office (Art.64) goes in the opposite direction. Initially proposed by the Parliament as an independent body, during the negotiation its role has been downsized. In the end, the AI office has been integrated within the Commission, as part of the administrative structure of the Directorate-General for Communication Networks, Content and Technology with limited scope and powers, staff provided by the DG CNECT and uncertain financial resources. So, a pivotal tassel to ensure accountability and achieve trustworthy AI will have to carry out its delicate function without being able to recruit personnel with an ad hoc expertise in the AI field, and limited autonomy in the exercise of its responsibility. 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.

References

AMNESTY INTERNATIONAL; ACCESS NOW. The Toronto Declaration: Protecting the right to equality and non-discrimination in machine learning systems. 2018. Available at: <https://www.torontodeclaration.org/>.

ASSOCIATION FOR COMPUTING MACHINERY US PUBLIC POLICY COUNCIL (USACM). Statement on Algorithmic Transparency and Accountability. 2017. Available at: <https://www.acm.org/binaries/content/assets/public-policy/2017_joint_statement_algorithms.pdf>.

BINNS, R. Algorithmic Accountability and Public Reason. Philosophy & Technology, v. 31, n. 4, p. 543–556, 2018. Available at: <https://doi.org/10.1007/s13347-017-0263-5>.

CATH, C.; FLORIDI, L. The Design of the Internet's Architecture by the Internet Engineering Task Force (IETF) and Human Rights. Sci Eng Ethics, v. 23, p. 449–468, 2017.

EUROPEAN COMMISSION. Communication 168 “Building Trust in Human-Centric Artificial Intelligence”. 2019.

EUROPEAN COMMISSION. Fostering a European Approach to Artificial Intelligence. 2020. Available at: <https://digital-strategy.ec.europa.eu/en/library/communication-fostering-european-approach-artificial-intelligence>.

EUROPEAN COMMISSION. White Paper on Artificial Intelligence. 2020. Available at: <https://commission.europa.eu/publications/white-paper-artificial-intelligence-european-approach-excellence-and-trust_en>.

FLORIDI, L. et al. AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations. Minds & Machines, v. 28, p. 689–707, 2018.

IEEE. Ethically Aligned Design First Edition. 2019. Available at: <https://standards.ieee.org/industry-connections/ec/ead1e-infographic/>.

LEWIS, D. et al. Global Challenges in the Standardization of Ethics for Trustworthy AI. Journal of ICT, v. 8, n. 2, p. 123–150, 2020.

MONTREAL DECLARATION FOR A RESPONSIBLE DEVELOPMENT OF ARTIFICIAL INTELLIGENCE. Available at: <https://recherche.umontreal.ca/english/strategic-initiatives/montreal-declaration-for-a-responsible-ai/>.

PALLADINO, N. A Digital Constitutionalism Framework for AI: security and fundamental rights in the Ai Act. Digital Politics, vol. 3, pp. 521-542, 2023, doi: 10.53227/113109.

PALLADINO, N.; SANTANIELLO, M. Legitimacy, power and inequalities in multistakeholder Internet governance. Cham: Palgrave McMillan, 2021. VAN KLYTON, A.; ARRIETA-PAREDES, M.-P.; PALLADINO, N.; SOOMAREE, A. Hegemonic practices in multistakeholder Internet governance: Participatory evangelism, quiet politics, and glorification of status quo at ICANN meetings. The Information Society, v. 39, n. 3, p. 141–157, 2023. Available at: <https://doi.org/10.1080/01972243.2023.2194295>.

RENDA, A. Artificial Intelligence, Ethics Governance and Policy Challenges. Brussels: CEPS, 2019. BOILER, G. Artificial Intelligence: The Great Disruptor. Washington, DC: The Aspen Institute, 2018.

RISSE, T. Transnational Governance and Legitimacy. In: BENZ, A. et al. (Eds.). Governance and Democracy Comparing National, European and International Experiences. New York: Routledge, 2006. p. 185.

RUSSELL, S.; NORVIG, P. Artificial Intelligence: A Modern Approach. Upper Saddle River, NJ: Prentice Hall, Pearson Education, 2003. (Prentice Hall Series in Artificial Intelligence).

SCHIFF, D. et al. Principles to Practices for Responsible AI: Closing the Gap (arXiv:2006.04707). arXiv, 2020. Available at: <http://arxiv.org/abs/2006.04707>.

SHILTON, K. “That’s Not an Architecture Problem!”: Techniques and Challenges for Practicing Anticipatory Technology Ethics. In: iConference 2015 Proceedings. iSchools, 2015.

SUZOR, N. et al. Human Rights by Design: The Responsibilities of Social Media Platforms to Address Gender-Based Violence Online. Policy & Internet, v. 11, p. 84–103, 2019.

TAMBIAMA, M. Artificial Intelligence Act Briefing (European Parliamentary Research Service). European Parliament, 2022.

THE SELF-REGULATORY CONVENTION FOR THE ARTIFICIAL INTELLIGENCE INDUSTRY (Draft for Comments). 2019. Available at: <https://www.secrss.com/articles/11099>.

TURNER, J. Robot Rules – Regulating Artificial Intelligence. London: Palgrave, 2019.
BROWN, I.; MARDSEN, C. Regulating Code. London: The MIT Press, 2013.
BROWNSWORD, R.; YEUNG, K. Regulating Technologies. Oxford: Hart Publishing, 2008.

UNESCO. Recommendation on the Ethics of Artificial Intelligence. 2022. Available at: <https://unesdoc.unesco.org/ark:/48223/pf0000381137>.

WIERINGA, M. 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, 2020.

Pre-print version of Belli L. Gaspar. W.B. The Quest for AI Sovereignty, Transparency and Accountability. Springer-Nature. (2025).

11. Promoting the Transparency of AI-Generated Inferences

Attamongkol 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. This practice allows companies to better understand individual consumer preferences. However, it often occurs secretly, without consumers being aware of it. Consumers' privacy may be compromised as they lack control over the accuracy, flow, and use of the generated inferential data. Although current data protection regulations, like the General Data Protection Regulation (GDPR), provide data subject rights, access to these inferences is not guaranteed. Businesses can deny inferences access requests by exploiting the broad scope of trade secrecy law, citing their interest in protecting such data as trade secrets. This Essay deems it essential to re-examine the scope and application of trade secrets law in this context. After providing a descriptive analysis of the underlying legal frameworks in the U.S. and EU that empower businesses to categorize consumer inferences as trade secrets, the Essay suggests 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.

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" (McKay, A., 2022).*

Using advanced algorithms to generate intimate inferences is no longer fictional. In today's personal-data-driven economy — either termed as "Surveillance Capitalism,"[1] "Informational Capitalism,"[2] or "Inference Economy"[3] — consumer data is not merely collected and used as is. Instead, consumer data is often processed and analysed using algorithms. The inferences drawn in the process will then be used by businesses, either to help formulate a more profitable strategy or to make critical decisions about the data subjects (Pasquale, F., 2015).

One example that has been around for quite some time is personalized behavioural advertisements. To show the right ads to the right customers, businesses typically use artificial intelligence (AI) tools to analyse data collected from consumers and generate a wide range of inferences about them, ranging from simple attributes to sensitive information such

[1] 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).

[2] 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.

[3] 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.

as physical and mental health, religious beliefs, and political views (Wachter, S., 2020). These inferences are then assigned to individuals in various forms, such as scores, tags, or categories (Solove, D. J., 2004). One grocery store chain in the U.S., Target, for instance, was revealed in 2012 to have secretly generated inferences about their 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 (Duhigg, C., 2012, February 16). Another notable example of inferences is in the infamous Facebook Cambridge Analytica scandal, where Facebook users were categorized into different political profiles based on their quiz answers and platform-monitored behavioural data, including "Like" clicks, and time spent on each post (Lewis, P. & Hilder P., 2018, March 23).

While businesses appear to get to know their consumers better through the generated inferential insights, consumers, on the other hand, know much less about the businesses and often lack awareness about the inferences concerning themselves (Marks, M., 2021). 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, the app 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." (Duportail, J., 2017, September 26). In 2018, Tinder maintained the same stance in response to a similar request from another user (Schmid, J., 2019, August 13). Similarly, back in 2011, Facebook rejected an access request made by privacy activist Max Schrems. Facebook's letter stated: "[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."[4]

The above instances, and many other, suggest that when the rights to trade secrets and the rights to personal data collide, the former seems to, by default, prevail. That is because disproving trade secret claims is always difficult. As the disapprover is not allowed to know the content of the secret to begin with, they are put in a less advantaged position (Kapczynski, A., 2022). For another reason, as will be further discussed in Part II, the seemingly limitless scope of trade secrets has been overlooked. And because data protection and privacy laws today prioritize trade secrets protection, data controllers can conveniently use their trade secrets interests as an excuse for not complying with data subjects' access requests ("DSARs"). Consequently, consumers are left with neither awareness nor control over inferences about them, despite how private and intimate such data could be. Therefore, this Essay aims to constrain—or at the very least, advocates for a re-examination of—the scope of trade secrets for the sake of data subjects' rights.

Existing literature often discusses the topic of transparency of AI algorithms, including how overly broad trade secret claims creatively made by coporations reduce AI transparency, undermining accountibility and due process (Kilic, B., 2024), the need for tranparent and accountable AI, especially in criniminal justice (Eaglin, J. M., 2017), and the idea of borrowing disclosure rules from security laws to promote algorithimic transparency (Lu, S., 2021), among others. This Essay, however, focuses on the transparency of AI products (i.e., inferences). It argues that inference transparency may be more beneficial for average consumers in making informed decisions about their rights (e.g., to edit or delete data

[4] 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).

concerning them) compared to algorithmic transparency, which may not fulfil this particular purpose due to inherent complexity of algorithmic models generally (Perl M. & Elkin-Koren N., 2017).

.

11.1. Cause of Inference Secrecy: The Underlying Legal Framework

The problem of inference opacity underscores a larger—and a global—issue: 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 existing regulations, including trade secrets law (Kapczynski, A., 2020). This Essay further establishes that alongside trade secrecy law, data privacy regulations also contribute to the issue, as the latter tends to prioritize businesses' interests in protecting trade secrets over data subjects' rights as the default stance.

11.1.1. 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 not known to the public, (ii) has commercial values thanks to its secrecy, and (iii) has been reasonably kept secret.[5] In the U.S., trade secrecy has long been recognized,[6] covering "all forms and types of financial, business, scientific, technical, economic, or engineering information" (Defend Trade Secrets Act of 2016, 18 U.S.C. § 1839, 3, 2018) as well as "a formula, pattern, compilation, program, device, method, technique, or process."[7]

As a result of the broad definition and protection of trade secrecy law, data-driven companies, such as Myriad Genetics and Google (Simon, B. M., & Sichelman T., 2017), claim not only their data analytics tools but also the data generated from those tools as trade secrets (Mattioli, M., 2014). Provided that the three factors (i.e., not publicly known, valuable thanks to secrecy, and kept confidential) are satisfied, personal data and inferences generated—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.[8]

11.1.2. 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.

---

[5] Agreement on Trade-Related Aspects of Intellectual Property (TRIPS) art. 39.

[6] See for example: Vickery v. Welch, 36 Mass. (19 Pick.) 523, 525 (1837)

[7] Uniform Trade Secrets Act § 1(4).

[8] 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 applicable in the United States as both the United States and the European Union are members of the TRIPS Agreement.

The EU General Data Protection Regulation (GDPR) acknowledges the right of access to personal data and relevant information[9]. 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 released a set of guidelines confirming that "observed," "derived," and "inferred" data are personal data that must be disclosed to a data subject upon request (The European Data Protection Board., 2022). Furthermore, in a case from that year, the European Court of Justice 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 (OT v. Vyriausioji tarnybinės etikos komisija, Case 184/20, 2022). As a result, legal experts interpreted that in most cases, generating sensitive inferences would trigger the requirement for obtaining explicit consent (Lomas, N., 2022). And ideally, by nature of the requirement, data subjects would be informed and able to retain some control.

However, since notice and consent generally occur before data processing, data subjects may not necessarily be informed about subsequently generated sensitive inferences, thereby leaving the right of access as the sole remedy. But because the GDPR limits the right of access by stating that the right "shall not adversely affect the rights and freedoms of others,"[10] including "trade secrets or intellectual property,"[11] data controllers may reject a DSAR if they believe that complying with it would compromise their trade secrets rights[12].

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[13]. What sets it apart is its explicit inclusion of inferences as one category of personal data[14], making it the first U.S. privacy law to do so (Blanke, J., 2020). Likewise, the Colorado law grants the right of access, encompassing inferences, (Colorado Privacy Act Rules 4 CCR-904-3, Rule 4.04, 2022) and go even further than the California law by providing a definition of "sensitive data inferences" with higher protection (Colorado Privacy Act Rules 4 CCR-904-3, Rule 2.02, 2022).

Nevertheless, both laws contain exceptions for trade secrets rights. The Colorado law states the exceptions explicitly[15], while the California law does so indirectly through provisions for future creation of exceptions concerning trade secrets and intellectual property rights.[16] Accordingly, the California State Attorney General noted in 2022 that businesses are not obligated to disclose inferences that qualify as trade secrets (Office of the Attorney General State of California., 2022). The prioritization of trade secrets of businesses over consumers' personal data rights is also evident in several other recent state privacy

[9] Directive 95/46/EC (General Data Protection Regulation) article 15

[10] Directive 95/46/EC (General Data Protection Regulation) article 15(4).

[11] Directive 95/46/EC (General Data Protection Regulation) recital 63.

[12] The European Data Protection Board. (n. 26) 49-50

[13] California Civil Code §1798.110(a)(1).

[14] California Civil Code §1798.140(o)(1)(K).

[15] 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").

[16] California Civil Code §1798.185.

legislations, including those of Tennessee, Iowa, Connecticut, Indiana, Montana, Oregon, Texas, and Utah.[17]

11.2. Data Subjects' Interest to Access Inferences about Themselves

11.2.1. Sophisticated Inference Generation

As the use of machine learning to process personal data becomes more common and cheaper, businesses can generate more inferences than before (Cofone, I. (2022). And these not just any inferences; "[AI] is increasingly able to derive the intimate from the available." (Calo, R., 2017). Advanced algorithms today make it possible to draw sensitive inferences from nearly all personal data, even the seemingly irrelevant ones (Solove, D., 2023). As researchers have pointed out: everything may not reveal everything; but in Big Data analytics, everything may reveal everything (Ohm, P., & Peppet, S., 2016).

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 as sexual orientation and political beliefs (Wachter, S., 2020).

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: Calls, texts, and apps usage	Big-Five personality traits: extraversion, agreeableness, conscientiousness, neuroticism, and openness to experience. (Chittaranjan, G., Blom, J., & Gatica-Perez, D., 2011)
Facebook likes	Ethnic affinity. (Angwin, J., & Parris, T., 2016)
	Sexual orientation, ethnicity, religious and political views, personality traits, intelligence, happiness, use of addictive substances, parental separation, age, and gender. (Kosinski, M., D. Stillwell, & T. Graepel., 2013)
Language on social media	Big-Five personality traits. (Park, G., Schwartz, H.A., Eichstaedt, J.C., Kern, M.L., Kosinski, M., Stillwell, D.J., Ungar, L.H., & Seligman, M.E., 2015)
Search query histories.	Age, gender, political and religious views. (Bi, B., Shokouhi, M., Kosinski, M. & Graepel, T., 2013)
Group photos	The importance of each person. (Mathialagan, C., Gallagher, A.C. & Batra, D., 2015)

[17] 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.

Rhythm of their typing patterns on a standard keyboard	Emotional state. (Epp, C., Lippold, M., & Mandryk, R.L., 2011)
Location tweets	Neighbourhoods of users, which then reveal average income, average housing cost, debt, and other demographic information, such as political views. (Liccardi, I., Abdul-Rahman, A., & Chen, M., 2016)

Furthermore, 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. The potential for intricate, sensitive, and intimate inference generation has also increased accordingly (Berrick, D., & Spivack, J., 2022). Some call the novel type of data that XR sensors can generate "Biometric Psychography." It is "a new quality of information that is comprised of the user's real identity combined with their reactions to particular stimuli, indicating what someone uniquely may think, like, and want." This includes: eye tracking and pupil response; facial scans; galvanic skin response, which shows emotion intensity; electroencephalography (EEG), which shows brain waves and state of mind; electromyography (EMG), which show muscle tension, capable of detecting truthfulness in a statement; and electrocardiography (ECG), which shows pulse and blood pressure (Heller, B., 2020).

Below are a few other examples of potential sensitive inferences that 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. (Bar-Zeev, A., 2019)
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. (Kröger, J. L., Lutz, O. H.-M., & Müller, F., 2020)
Behavioral data (e.g., reaction time, voice, vision, fitness)	Age and disabilities. (Nair, V., Garrido G. M., & Song, D., 2022)
Device data (e.g., CPU power, resolution, tracking and refresh rates)	Wealth (Nair, V., Garrido, G. M., & Song, D., 2022)
Geospatial data (e.g., height, left & right arms, room)	Wingspan, gender, room area (Nair, V., Garrido G. M., & Song, D., 2022)

11.2.2. The Case for Accessing Inferences

Despite the potential benefits that inferences can offer (e.g., positive nudges (Möhlmann, M., Apr. 22, 2021), enhanced service experiences[18], and facilitated markets), (Kim, P.T., 2020) sensitive inference generation implicates a series of interrelated risks, namely discrimination, manipulation, privacy loss, and potential harm from inaccuracies.

As scholars have argued, "personalization is another word for *discrimination*." (Ohm P., & Peppet S., 2016). Notably, many online publishers are known for allowing and supporting businesses to display advertisements based on users' protected-class profiles, which can be more easily identified over time thanks to more extensive data collection practices and improved Big Data analytics (Dobkin, A., 2018). Consider as examples when black people's names searches display more advertisements by arrest search websites (Wachter, S., 2020), job search engines show more STEM positions to male job seekers compared to females (Alegre S., 2021), and online tutoring providers suggest double-priced SAT courses to Asian teens. (Angwin, J., et al., 2016). Trained on data already embedded with pre-existing social bias, inference-based personalization often reinforce and worsen these patterns of discrimination (Kim, P.T., 2020).

Second, uses of inference raise concerns about *manipulation* (Sunstein, C.R., 2015). For example, consider Facebook's inference-based feeds which have allegedly altered not only users' emotional states (Kramer, A. D. I., et al., 2014) but also their political stances (Kim, P.T., 2020). Online platforms also use inference-led personalization as a psychological tool to keep users on screen for the longest period possible, thereby maximizing their profits in the attention economy through the increased number of ads displayed. To Susie Alegre, these practices can be seen as manipulation of thoughts and opinions, which goes against the internationally protected "Right to Freedom of Thought" that is fundamentally important in the digital age (Alegre S., 2021).

Third, inference generation can give rise to concerns about *privacy loss*. Commentators argue that privacy grants individuals with the right to determine their own identities and beliefs (Richards N., 2011). Through a pattern of "Like" clicks on a platform, the platform can potentially uncover not only a user's preference but also, with AI assistance, their deeper, intimate thoughts and opinions regarding certain beliefs, objects, and fantasies. From this lens, inferences and predictions generated by businesses thus take away some of the consumers' ability to privately identify themselves. It also takes away "an ability to decide the extent to which one's inner most thoughts, desires, and domestic relations are shared," which is what the "Right to Privacy" means to notable scholars like Samuel Warren and Louis Brandies and a critical component of "Intimate Privacy" for Danielle Citron (Citron, D. K., 2022).

Fourth, large-scale inference generation often leads to *inaccuracies*, (Cyphers, B., & Gebart, G., 2019) with corrections rarely possible (Szymielewicz, K., 2019). 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 (Blanke, J., 2020, 84-85). As a consequence, these false inferences could potentially lead to an unforeseeable harm for the individual (Cofone, I., 2022). Even if told that the inference is

[18] 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).

false, the individual identified by the inference is still subject to risk of being perceived as shameful or treated with bias from others due to a common misinterpretation of correlative and causal associations. For example, "[t]he man whose name was associated with fraud on Google may or may not be the cause of that association, but we certainly read it that way." (Diakopoulos, N., 2018). This note suggests that people tend to assume that individuals identified with negative inferences must have done something in the past to cause the algorithm to generate these inferences, even when that is not necessarily the case, considering how algorithms work.

When data subjects struggle to predict the extent of information held by a data controller and its potential repercussions due to trade secrets, they lack the control over their personal data and the ability to reduce potential risks. In data protection terms, the inability to access inferences makes the data subject lack the knowledge needed to effectively use other data protection remedies—such as correcting or deleting data about them. Commentators contend that the right of access, as a cornerstone for other rights (Ausloos, J., Veale, M., & Mahieu R., 2020), therefore, should apply to inferences which the data subjects are typically less aware of (Shah, S., 2019). The California State Attorney General is also in alignment with this view, as it commented that "inferences appear to be at the heart of the problem that the CCPA seeks to address." (Office of the Attorney General State of California., 2022) Nonetheless, the trade secrets issue is often left untouched.

Wachter and Mittelstadt argue for the "Right to Reasonable Inferences" to regulate high-risk inferences (i.e., causing potential reputation damage) that are not verifiable in nature (Wachter, S. & Mittelstadt, B., 2019, 494-591). 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 (Ibid, 606-610). Thus, without challenging the application of trade secrets law, the inference opacity problem would remain.

11.3. Challenging Trade Secrecy for Inference Transparency

Therefore, this Essay takes the approach of questioning the law of trade secrecy itself, which appears to be the root cause of the problem. Below, it examines the literature on U.S. trade secrets law, offering a framework that data protection agencies (or consumer protection agencies, depending on jurisdictions) may follow. This approach is preferable to judicial remedies, such as private actions, as challenging trade secret claims in court is typically difficult, costly, and time-consuming (Kilic, B., 2024, p.13).

The framework argues that when companies reject DSARs by claiming trade secrets protection, they make two assertations: 1) that inferences generated by them are always eligible for protection as trade secrets; and 2) that trade secrets law grants property-like exclusive rights that can be exercised against the data subjects. As will be explained further, the two assumptions are contestable and should be contested. Frequently, information claimed as proprietary does not meet the established standards for legal protection as trade secrets (Kilic B., 2024). Data protection agencies may address these challenges in comments or guidelines to facilitate data subjects' access to their inferences.

11.3.1. Examine which particular inferences can qualify for protection as trade secrets

When a business says that they cannot comply with DSARs related to inferences, it assumes that the inferences it generates qualify as trade secrets. While it is generally understood that any information that is *secret*, *valuable*, and *kept secret* can be protected as a trade secret, there may exist specific legal definitions or nuances that the authorities may highlight to prevent potential abuse claims.

In the U.S., the *secrecy* requirement used to entail only the information that is not readily ascertainable, determined by the level of effort and difficulty required to obtain such information (Dole, R. F., 2016); the *valuable* requirement applies only to information that gains independent economic value from not being generally known to others (Johnson, E., 2010); and the subject matter ought to be continuously used in commerce—*keeping it secret* is not enough (Lemley, M. A. & Hrdy, C. A., 2021).

By adhering to the definitions above, one will find that not every inference can be protected as a trade secret. First, some inferences that are more obvious than others would unlikely pass the *secrecy* prong. These may include inferences that are readily apparent, such as gender, race, and age range inferred from a photo, and are thus likely to be already known by multiple parties, including competitors—and, of course, data subjects themselves. Second, inferences that can provide commercial benefits independently of their secrecy would not satisfy the *valuable* requirement. And that might apply to most inferences. To clarify, inferences do not necessarily need to be kept secret for businesses to effectively deliver personalized ads or services. For example, Company A having the same inferential insights about a consumer as Company B should have minimal, if not zero, impact on Company A's ability to continue offering targeted ads. Lastly, abandoned inferences (i.e., inferences generated or acquired but never used) would lose their protection as trade secrets. The inferences that fail to meet any of the three requirements should therefore be disclosed to the data subjects upon request.

11.3.2. Examine the nature and scope of trade secrets rights

When businesses refuse to disclose inferences through DSARs, they are also assuming that trade secrets law grant them property-like exclusive rights that can be exercised against any party, including data subjects. While a contemporary interpretation of trade secrecy law may appear to align with the concept of trade secrets being intellectual property (IP), its historical roots may suggest otherwise. Therefore, the authorities should critically assess the interpretation of trade secrets as IP and interpret trade secrets law in their respective jurisdictions with the interests of data subjects in mind when issuing future DSARs guidelines[19].

For instance, Kapczynski suggests that in the U.S., trade secrets law originally resembled torts law rather than the "right of property in the idea" (Kapczynski, A., 2022).

[19] This Essay holds the opinion that data protection agencies can comment on the scope of trade secrets protection when it conflicts with data subjects' rights. The discussion of trade secret law receives scant attention in academic literature and is rarely addressed by official authorities. 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*, 548 ("So where Congress authorizes the PTO to comment or advice on intellectual property, in theory the [Defend Trade Secrets Act] is not included").

Over time, it became part of unfair competition law, labour law, and contracts law. The assertion of trade secrets rights was thus 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, places a greater emphasis on secrecy over wrongful conduct. Within the context of DSARs, this approach is evident in businesses asserting their trade secrets interests against data subjects, even when no wrongful conduct has occurred. In a broader context, the concept of trade secrets as IP has allowed U.S. businesses to keep any information away from public scrutiny as they wish. Therefore, scholars like Kapczynski highlight the importance of adhering to the historical roots of trade secrets law to limit corporate power over information.

11.3.3. Require data controllers to document access requests reasons in detail.

Additionally, the agencies are advised to require data controllers to document how they handle DSARs in detail. This Essay 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 the respective jurisdiction—which must be made available for authority's audit. Note that the GDPR at Article 30 does not require data controllers to document DSARs refusal grounds in its ROPA. However, this practice is suggested by the data protection authority in Singapore[20], and this Essay proposes that other jurisdictions follow this approach to address DSAR denial abuses.

11.4. 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 Essay presents a framework for data protection authorities in any jurisdictions facing a similar issue to adopt. Specifically, it prompts the authorities to critically assess the subject matters of trade secrets and the scope of rights in their jurisdiction and require data controllers to document in detail the reasons for denying DSARs pertaining inferences.

Although outside scope of this Essay, further studies on the subsequent remedies for data subjects upon becoming aware of inferences about themselves are crucial. Further discussions 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[21]. 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

[20] *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").

[21] 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)

generate these inferences remain unaltered, there exists a possibility that that the "algorithmic shadow" could regenerate the same inferences even after deletion (Li, T. C., 2022).

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.

References

AGREEMENT ON TRADE-RELATED ASPECTS OF INTELLECTUAL PROPERTY (TRIPS).

ALEGRE, S. Protecting freedom of thought in the digital age. Center of International Governance Innovation, n. 165, p. 3-4, 2021. Available at: <https://www.cigionline.org/static/documents/PB_no.165.pdf>.

ALLEN, A. L. Coercing privacy. William & Mary Law Review, v. 40, p. 738, 1999.

ANGWIN, J. et al. When Algorithms Decide What You Pay. 2016. Available at: <https://www.propublica.org/article/breaking-the-black-box-when-algorithms-decide-what-you-pay#:~:text=Charging%20different%20prices%20to%20different,Princeton%20Review%20than%20non%2DAsians>.

ANGWIN, J.; PARRIS, T. Facebook lets advertisers exclude users by race. 2016. Available at: <https://www.propublica.org/article/facebook-lets-advertisers-exclude-users-by-race>.

ARTICLE 29 DATA PROTECTION WORKING PARTY. Guidelines on automated individual decision-making and profiling. 2016. p. 18.

AUSLOOS, J.; VEALE, M.; MAHIEU, R. Getting data subject rights right. Journal of Intellectual Property, Information Technology, and Electronic Commerce Law, 2020. Available at: <https://doi.org/10.31228/osf.io/e2thg>.

BAR-ZEEV, A. The eyes are the prize: Eye-tracking technology is advertising's holy grail. 28 May 2019. Available at: <https://www.vice.com/en/article/bj9ygv/the-eyes-are-the-prize-eye-tracking-technology-is-advertisings-holy-grail>.

BERRICK, D.; SPIVACK, J. Understanding extended reality technology & data flows: Privacy and data protection risks and mitigation strategies. 17 Nov. 2022. Available at: <https://fpf.org/blog/understanding-extended-reality-technology-data-flows-privacy-and-data-protection-risks-and-mitigation-strategies/>.

BI, B.; SHOKOUHI, M.; KOSINSKI, M.; GRAEPEL, T. Inferring the demographics of search users: Social data meets search queries. In: Proceedings of the 22nd international conference on World Wide Web, 2013. p. 131-140. Available at: <https://doi.org/10.1145/2488388.2488401>.

BLANKE, J. Protection for 'inferences drawn:' A comparison between the General Data Protection Rule and the California Consumer Privacy Act. Global Privacy Law Review, v. 2, p. 92, 2020.

CALIFORNIA CIVIL CODE.

CALO, R. Artificial intelligence policy: A primer and roadmap. University of California Davis Law Review, v. 51, p. 421, 2017.

CHITTARANJAN, G.; BLOM, J.; GATICA-PEREZ, D. Who's who with big-five: Analyzing and classifying personality traits with smartphones. In: Wearable Computers

(ISWC), 2011 15th Annual International Symposium, 2011. p. 29-36. Available at: <https://doi.org/10.1109/ISWC.2011.29>.

CITRON, D. K. The fight for privacy: Protecting dignity, identity, and love in the digital age. W. W. Norton & Company, 2022. p. xiii.

COFONE, I. Privacy standing. Illinois Law Review, p. 1384, 2022.

COHEN, J. E. Between truth and power. Oxford University Press, 2019. p. 5-6.

COLORADO PRIVACY ACT RULES. 4 CCR-904-3, 29 Sep. 2022.

CYPHERS, B.; GEBART, G. Behind the one-way mirror: A deep dive into the technology of corporate surveillance. 2019. Available at: <https://www.eff.org/files/2019/12/11/behind_the_one-way_mirror-a_deep_dive_into_the_technology_of_corporate_surveillance.pdf>.

DEFEND TRADE SECRETS ACT OF 2016, 18 U.S.C. § 1839(3) (2018).

DIAKOPOULOS, N. Accountability in algorithmic decision making. Communications of the ACM, v. 59, n. 2, p. 58, 2018. Available at: <https://doi.org/10.1145/2844110>.

DIRECTIVE 95/46/EC (General Data Protection Regulation).

DOBKIN, A. Information fiduciaries in practice: Data privacy and user expectation. Berkeley Technology Law Journal, v. 33, p. 26-27, 2018.

DOLE, R. F. The contours of American trade secret law: What is and what isn't protectable as a trade secret. SMU Science and Technology Law Review, v. 19, p. 99, 2016.

DUHIGG, C. How companies learn your secrets. The New York Times, 16 Feb. 2012. Available at: <https://www.nytimes.com/2012/02/19/magazine/shopping-habits.html?searchResultPosition=1>.

DUPORTAIL, J. I asked Tinder for my data. It sent me 800 pages of my deepest, darkest secrets. The Guardian, 26 Sep. 2017. Available at: <https://www.theguardian.com/technology/2017/sep/26/tinder-personal-data-dating-app-messages-hacked-sold>.

EAGLIN, J. M. Constructing recidivism risk. Emory Law Journal, v. 67, p. 59, 2017.

EPP, C.; LIPPOLD, M.; MANDRYK, R. L. Identifying emotional states using keystroke dynamics. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, 2011. p. 715-724. Available at: <https://doi.org/10.1145/1978942.1979046>.

FACEBOOK. Copy of the letter Facebook sent to Schrems. Available at: <http://www.europe-v-facebook.org/FB_E-Mails_28_9_1.pdf>. Acesso em: 6 Jul. 2023.

For each state bill, see IAPP U.S. State Privacy Legislation Tracker 2023. Available at: <https://iapp.org/media/pdf/resource_center/State_Comp_Privacy_Law_Chart.pdf>.

GOLDMAN, E. The Defend Trade Secrets Act isn't an "intellectual property" law. Santa Clara High Tech, v. 33, 2016.

HELLER, B. Watching androids dream of electric sheep: Immersive technology, biometric psychography, and the Law. Vanderbilt Journal of Entertainment & Technology, v. 23, p. 27-29, 2020.

JOHNSON, E. Trade secret subject matter. Hamline Law Review, v. 33, p. 556, 2010.

KAPCZYNSKI, A. The law of informational capitalism. Yale Law Journal, v. 129, p. 1501, 2020.

KAPCZYNSKI, A. The public history of trade secrets. University of California Davis Law Review, v. 55, p. 1384-1442, 2022.

KILIC, B. Into uncharted waters: trade secrets law in the AI era. CIGI Papers, no. 295, p 10-12, 2024. Available at: <https://www.cigionline.org/static/documents/no.295.pdf>

KIM, P. T. Manipulating opportunities. Virginia Law Review, v. 106, p. 868, 2020.

KOSINSKI, M.; STILLWELL, D.; GRAEPEL, T. Private traits and attributes are predictable from digital records of human behaviour. Proceedings of the National Academy of Sciences, v. 110, n. 15, p. 5802-5805, 2013. Available at: <https://doi.org/10.1073/pnas.1218772110>.

KRAMER, A. D. I. et al. Experimental evidence of massive-scale emotional contagion through social networks. Proceedings of the National Academy of Sciences of the United States of America, v. 111, n. 24, p. 8788–8790, 2014. Available at: <www.pnas.org/content/111/24/8788>.

KRÖGER, J. L.; LUTZ, O. H.-M.; MÜLLER, F. 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, p. 226-241, 2020. Available at: <https://doi.org/10.1007/978-3-030-42504-3_15>.

LEMLEY, M. A.; HRDY, C. A. Abandoning trade secrets. Stanford Law Review, v. 74, p. 4, 2021.

LEWIS, P.; HILDER, P. Leaked: Cambridge Analytica's blueprint for Trump victory. The Guardian, 23 Mar. 2018. Available at: <https://www.theguardian.com/uk-news/2018/mar/23/leaked-cambridge-analyticas-blueprint-for-trump-victory>.

LI, T. C. Algorithmic destruction. Southern Methodist University Law Review, v. 75, p. 482, 2022.

LICCARDI, I.; ABDUL-RAHMAN, A.; CHEN, M. 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, 2016. p. 1-12. Available at: <https://doi.org/10.1145/2858036.2858272>.

LOMAS, N. Sensitive data ruling by Europe's top court could force broad privacy reboot. 2 Aug. 2022. Available at: <https://techcrunch.com/2022/08/02/cjeu-sensitive-data-case/>.

LU, S. Algorithmic opacity, private accountability, and corporate social disclosure in the age of artificial intelligence. Vanderbilt Law Review, v. 23, p. 99, 2021.

MARKS, M. Emergent medical data: Health information inferred by artificial intelligence. U.C. Irvine Law Review, v. 11, 2021.

MATHIALAGAN, C.; GALLAGHER, A. C.; BATRA, D. Vip: Finding important people in images. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015. p. 4858-4866. Available at: <https://research.google/pubs/pub43844/>.

MATTIOLI, M. Disclosing big data. Minnesota Law Review, v. 99, p. 556, 2014.

MCKAY, A. Don't Look Up [Motion Picture]. 2022.

MÖHLMANN, M. Algorithmic nudges don't have to be unethical. Harvard Business Review, 22 Apr. 2021. Available at: <https://hbr.org/2021/04/algorithmic-nudges-dont-have-to-be-unethical>.

NAIR, V.; GONZALO MUNILLA GARRIDO; SONG, D. Exploring the unprecedented privacy risks of the Metaverse. 23rd Privacy Enhancing Technologies Symposium, p. 238-256, 2022. Available at: <https://doi.org/10.48550/arxiv.2207.13176>.

OFFICE OF THE ATTORNEY GENERAL STATE OF CALIFORNIA. Opinion No. 20-303, p. 14-15, 10 Mar. 2022.

OFFICE OF THE ATTORNEY GENERAL STATE OF CALIFORNIA. Opinion No. 20-303, p. 13, 10 Mar. 2022. Available at: <https://oag.ca.gov/system/files/opinions/pdfs/20-303.pdf>.

OHM, P.; PEPPET, S. What if everything reveals everything?. In: SUGIMOTO, C.; EKBIA, H.; MATTIOLI, M. (Eds.). Big Data Is Not a Monolith. MIT Press, 2016. p. 47.

OT v. Vyriausioji tarnybinės etikos komisija. Case 184/20 (CJEU Aug. 1, 2022).

PARK, G. et al. Automatic personality assessment through social media language. Journal of Personality and Social Psychology, v. 108, n. 6, p. 934, 2015. Available at: <https://doi.org/10.1037/pspp0000020>.

PASQUALE, F. The black box society. Harvard University Press, 2015. p. 25-32.

PERL, M.; ELKIN-KOREN, N. Black box tinkering: Beyond disclosure in algorithmic enforcement. Florida Law Review, v. 69, p. 186-90, 2017.

PERSONAL DATA PROTECTION COMMITTEE SINGAPORE. Guide to handling access requests. 9 Jun. 2016. Available at: <https://www.pdpc.gov.sg/-/media/Files/PDPC/PDF-Files/Other-Guides/guide-to-handling-access-requests-v1-0-(090616).pdf>.

PERSONALIZATION-PRIVACY PARADOX: WHY SOLVING IT MATTERS NOW. 4 Jul. 2022. Available at: <https://www.cdotrends.com/story/16573/personalization-privacy-paradox-why-solving-it-matters-now?refresh=auto>.

RICHARDS, N. Why privacy matters. Oxford University Press, 2011.

SCHMID, J. My GDPR complaint against tinder [web blog comment]. 13 Aug. 2019. Available at: <https://medium.com/personaldata-io/my-gdpr-complaint-against-mtch-technology-services-139087d3de8a>.

SHAH, S. This lawyer believes GDPR is failing to protect you: Here's what we should change. 30 Jan. 2019. Available at: <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>.

SIMON, B. M.; SICHELMAN, T. Data-generating patents. Northwestern University Law Review, v. 111, p. 377, 2017.

SOLOVE, D. Data is what data does: Regulating use, harm, and risk instead of sensitive data. Northwestern University Law Review, v. 118, 2023. Available at: <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4322198>.

SOLOVE, D. J. The digital person: Technology and privacy in the information age. New York University Press, 2004. p. 46.

SOLOW-NIEDERMAN, A. Information privacy and the inference economy. Northwestern University Law Review, v. 117, p. 361, 2022.

SOLOW-NIEDERMAN, A.; COFONE, I. Privacy standing. Illinois Law Review, p. 1384, 2022. UbertI, D. Come the metaverse, can privacy exist? Wall Street Journal, 4 Jan. 2022. Available at: <https://www-wsj-com.cdn.ampproject.org/c/s/www.wsj.com/amp/articles/come-the-metaverse-can-privacy-exist-11641292206>.

SUNSTEIN, C. R. Fifty shades of manipulation. **Journal of Marketing Behavior**, v. 1, p. 216-218, 2015.

SZYMIELEWICZ, K. Your digital identity has three layers, and you can only protect one of them. 25 Jan. 2019. Available at: <https://qz.com/1525661/your-digital-identity-has-three-layers-and-you-can-only-protect-one-of-them/>.

TENE, O.; POLONETSKY, J. Big data for all: Privacy and user control in the age of analytics. **Northwestern Journal of Technology & Intellectual Property**, v. 11, n. 5, p. 270, 2013.

THE EUROPEAN DATA PROTECTION BOARD. 2022. p. 49-50.

THE EUROPEAN DATA PROTECTION BOARD. Guidelines 01/2022 on data subject rights - right of access. 1 Feb. 2022. Available at: <https://edpb.europa.eu/our-work-tools/documents/public-consultations/2022/guidelines-012022-data-subject-rights-right_en>.

UNIFORM TRADE SECRETS ACT.

VICKERY v. Welch, 36 Mass. (19 Pick.) 523, 525 (1837).

WACHTER, S. 2020. p. 376-378.

WACHTER, S. Affinity profiling and discrimination by association in online behavioural advertising. **Berkeley Tech Law Journal**, v. 35, p. 376-377, 2020.

WACHTER, S.; MITTELSTADT, B. 2019. p. 494-591.

WACHTER, S.; MITTELSTADT, B. A right to reasonable inferences: Re-thinking data protection law in the age of big data and AI. **Columbia Business Law Review**, v. 2019, p. 607, 2019.

WACHTER, S.; MITTELSTADT, B.; FLORIDI, L. Why a right to explanation of automated decision-making does not exist in the general data protection regulation. International Data Privacy Law, v. 7, n. 2, p. 85-89, 2017. Available at: <https://doi.org/10.1093/idpl/ipx005>.

WEISS, B. Trump-linked firm Cambridge Analytica collected personal information from 50 million Facebook users without permission. Business Insider, 18 Mar. 2018. Available at: <https://www.businessinsider.com/cambridge-analytica-trump-firm-facebook-data-50-million-users-2018-3>.

ZUBOFF, S. The age of surveillance capitalism: The fight for a human future at the new frontier of power. 2019.

Pre-print version of Belli L. Gaspar. W.B. The Quest for AI Sovereignty, Transparency and Accountability. Springer-Nature. (2025).

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[1]

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, 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.

[1] 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 pape

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)[2] 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"* *(Austin, W., Park, C., & Goble, E., 2008)* of the 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?

[2] NATO Emerging and disruptive technologies, Retrieved August 2023, from https://www.nato.int/cps/en/natohq/topics_184303.htm.

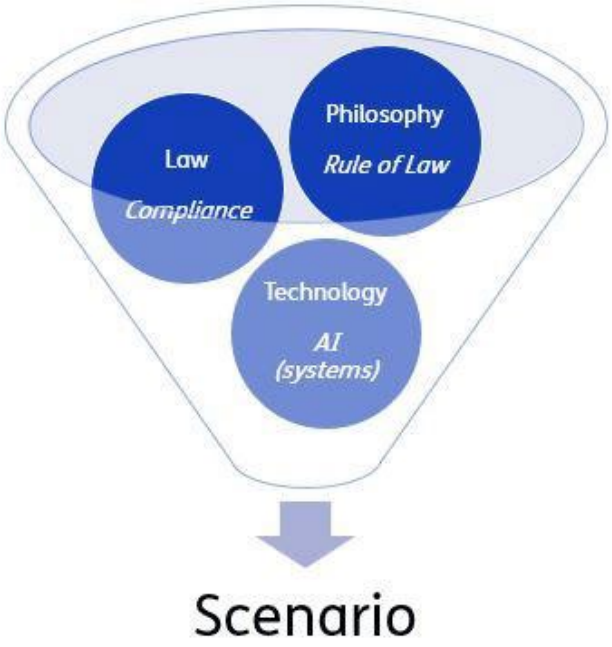


Figure 1: Interdisciplinary approach in the context of scenarios

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

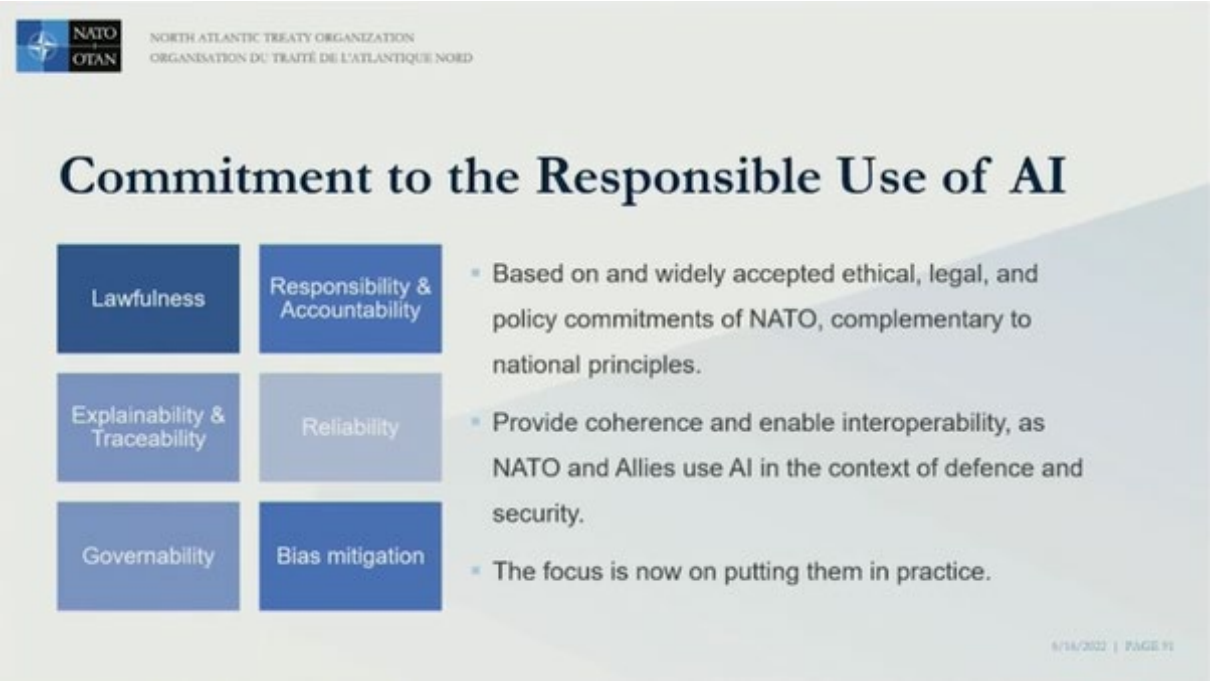
In this section two NATO principles of responsible use (*lawfulness* and *governability*) are introduced and contextualised via the Rule of Law. What is responsible AI, and what is its relevance with respect to Rule of Law tenets and mechanisms? Responsible AI starts at the acquisition process, or – 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?

Figure 2: Illustration of the six principles of responsible use of AI: lawfulness; responsibility & accountability; explainability; reliability; governability; bias mitigation. (NATO, 2022)

The Rule of Law[3] 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." (United Nations., n.d.),[4]

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

[3] 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."*

[4] See also: Article 2 of the Treaty on European Union.

may lead to erosion of NATO's principle of responsible use: *lawfulness*, and consequently also of the tenets of the Rule of Law.

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

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 (Zoe Stanley Lockman, E. H., 2021), OECD (OECD.AI Policy Observatory., 2023), the European Commission (European Commission., 2021), the US Department of Defense (Board, D. I., 2019), and the UK Ministry of Defence (Ministry of Defence United Kingdom., 2022). 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 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." (International Court of Justice, 1996)* 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]."*[5]

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. 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-*

[5] 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))

> *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." (Hofman, J. M., Sharma, A., & Watts, D. J., 2017)*

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."*

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[6]. 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[7]. Another example of military usage of AI is swarm

[6] 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.

[7] 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.

intelligence: artificial intelligence, often used in UAS, acting in a coordinated manner without a central control unit.[8]

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. 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[9] 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[10] (DARB) which has set the goal of developing a toolkit for Responsible AI Certification Standards building on experiences gained from operational use cases. This 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

[8] Sentient Digital. MILITARY DRONE SWARM INTELLIGENCE EXPLAINED, Retrieved from September 2023, from https://sdi.ai/blog/military-drone-swarm-intelligence-explained/.

[9] Frameworks of: NATO, OECD, European Commission, the US Department of Defense, and the UK Ministry of Defence.

[10] NATO's Data and Artificial Intelligence Review Board, Retrieved August 2023, from https://www.nato.int/cps/en/natohq/official_texts_208374.htm.

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) [11] 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".*[12]

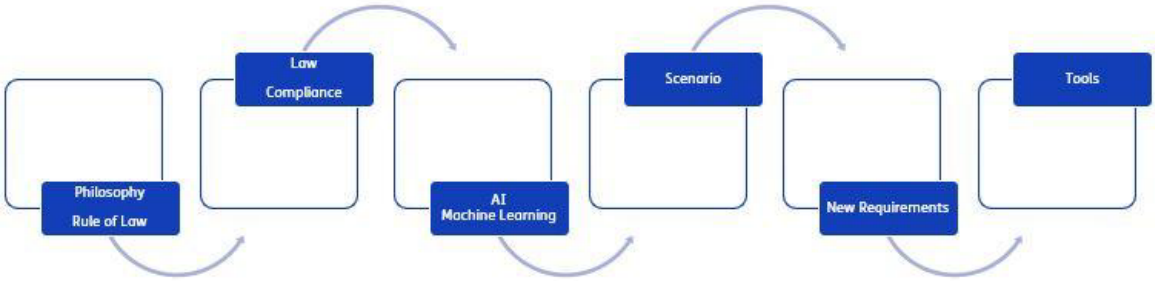


Figure 4: How to use scenarios to find new requirements which can become tools?

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.

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.

[11] The NATO principles of responsible use include: lawfulness; responsibility and accountability; explainability and traceability; reliability; governability; and bias mitigation.

[12] NATO's Data and Artificial Intelligence Review Board, Retrieved August 2023, from https://www.nato.int/cps/en/natohq/official_texts_208374.htm.

An electro-optic camera system can be enhanced with an AI application in order to automate the distinction of small Unmanned Aircraft Systems (sUAS)[13] 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 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."*[14] 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."*[15] 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."*[16]*)* 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.

---

[13] sUAS refers to small UAS, which is a subcategory of NATO Class I UAS with a weight below 15 kgs.

[14] NATO's Data and Artificial Intelligence Review Board, Retrieved August 2023, from https://www.nato.int/cps/en/natohq/official_texts_208374.htm.

[15] NATO's Data and Artificial Intelligence Review Board, Retrieved August 2023, from https://www.nato.int/cps/en/natohq/official_texts_208374.htm.

[16] NATO's Data and Artificial Intelligence Review Board, Retrieved August 2023, from https://www.nato.int/cps/en/natohq/official_texts_208374.htm.

Consequently, the (AI) technology may violate the mechanisms of the Rule of Law i.e. *"procedural and legal transparency." (United Nations., n.d.)* Since accountability questions cannot be addressed.

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 cannot 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.

References

Austin, W.; Park, C.; Goble, E. From interdisciplinary to transdisciplinary research: A case study. Qualitative Health Research, v. 18, n. 4, p. 557-564, 2008.

BOARD, D. I. AI principles: recommendations on the ethical use of artificial intelligence by the department of defense: supporting document. United States Department of Defense, 2019.

European Commission. Ethics Guidelines for Trustworthy AI. 2021. Available at: <https://ec.europa.eu/futurium/en/ai-alliance-consultation.1.html>. Acesso em: ago. 2023.

European Parliament; Council of the European Union. 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)).

Hofman, J. M.; Sharma, A.; Watts, D. J. Prediction and explanation in social systems. Science, v. 355, n. 6324, p. 486-488, 2017.

INTERNATIONAL COURT OF JUSTICE. Legality Of The Threat Or Use Of Nuclear Weapons Advisory Opinion Of 8 Jul 1996, para. 78.

Ministry of Defence United Kingdom. Policy paper Ambitious, safe, responsible: our approach to the delivery of AI enabled capability in Defence. 2022. Available at: <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>. Acesso em: ago. 2023.

National Defense Magazine. Ukraine A Living Lab for AI Warfare. Available at: <https://www.nationaldefensemagazine.org/articles/2023/3/24/ukraine-a-living-lab-for-ai-warfare>. Acesso em: set. 2023.

NATO. Emerging and disruptive technologies. Available at: <https://www.nato.int/cps/en/natohq/topics_184303.htm>. Acesso em: ago. 2023.

NATO. Future conflicts may be won or lost by AI. Apresentado por Nikos Loutas, Head of Data and Artificial Intelligence Body, NATO, in: Tech Informed. Available at: <https://techinformed.com/nato-future-conflicts-may-be-won-or-lost-by-ai/>. Acesso em: set. 2023.

NATO. NATO's Data and Artificial Intelligence Review Board. Available at: <https://www.nato.int/cps/en/natohq/official_texts_208374.htm>. Acesso em: ago. 2023.

OECD.AI Policy Observatory. OECD AI Principles overview. 2023. Available at: <https://oecd.ai/en/ai-principles>. Acesso em: ago. 2023.

Sentient Digital. Military Drone Swarm Intelligence Explained. Available at: <https://sdi.ai/blog/military-drone-swarm-intelligence-explained/>. Acesso em: set. 2023.

Stanley Lockman, Z.; H., E. An Artificial Intelligence Strategy for NATO. NATO Review, 25 Oct. 2021. Available at: <https://www.nato.int/docu/review/articles/2021/10/25/an-artificial-intelligence-strategy-for-nato/index.html>. Acesso em: ago. 2023.

sUAS refers to small UAS, which is a subcategory of NATO Class I UAS with a weight below 15 kgs.

United Nations. What is the Rule of Law. Available at: <https://www.un.org/ruleoflaw/what-is-the-rule-of-law/#:~:text=It%20requires%20measures%20to%20ensure,and%20procedural%20and%20legal%20transparency>. Acesso em: ago. 2023.

United Nations. What is the Rule of Law. Available at: <https://www.un.org/ruleoflaw/what-is-the-rule-of-law/#:~:text=It%20requires%20measures%20to%20ensure,and%20procedural%20and%20legal%20transparency>. Acesso em: ago. 2023.

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

Pre-print version of Belli L. Gaspar. W.B. The Quest for AI Sovereignty, Transparency and Accountability. Springer-Nature. (2025).

13. Iterating AI Accountability in the Chinese Model AI Law: From Fragmentation to Meaningful Generalization

Wayne Wei Wang, PhD Candidate, Faculty of Law, University of Hong Kong & Non-Resident Fellow 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

This chapter, initially drafted as and developed upon a policy paper for the DC-DAIG, examines the divergent conceptualizations of Artificial Intelligence (AI) accountability among diverse stakeholders and actors, laying the groundwork for a nuanced discourse on the inherent ambiguities and practical challenges in establishing normative frameworks for AI governance. In particular, the chapter critically assesses current global regulatory efforts targeting Generative AI, especially as they intersect with data protection legislation. It reveals how these varied regulatory approaches have coalesced into an interim, composite framework in the absence of comprehensive, formalized statutes. Using China as a focal jurisdictional case study, this analysis tracks the evolving regulatory landscape regarding ethical principles, content security, and data protection standards. The chapter juxtaposes the binding "Interim Measures for Generative AI" with the proposed, non-binding "Chinese Model AI Law," introduced by leading Chinese academics in 2023 and iteratively revised in 2024, to illustrate contrasts and tensions in regulatory theory and practice. Ultimately, the chapter contends that achieving operational AI accountability demands a robust institutional framework that clearly defines specific obligations and liabilities, such as data retention, disclosure mandates, and mechanisms for cross-border cooperation. It further argues for a carefully calibrated balance between regulatory flexibility and predictability, essential to fostering an adaptable, enforceable, and practicable accountability framework for AI.

Introduction

The prospect of holding artificial intelligence systems accountable, once regarded as a daunting challenge, remains inextricably linked to the complex interplay of its constituent dimensions, including explainability, accuracy, reliability, and robustness. A notable example is the European Union's AI Act, particularly its Article 1 and Recital 1, which spurred debates about the tension between fostering innovation and imposing regulatory constraints. These discussions often centred on the nuanced distinctions between explainability and interpretability (Grady, 2022). While some AI models may resist complete interpretability, a variety of model-agnostic methodologies have been developed to shed light on their inner workings, adhering to established principles of machine-learning explanation (Du et al., 2019). However, in critical applications—especially those involving adversarial settings—post-hoc explanation mechanisms designed to enhance explainability frequently fail to achieve optimal functionality (Bordt et al., 2022).

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). In contrast to human

agents, AI entities lack moral agency and sentience, complicating the attribution of culpability or their amenability to account (Coeckelbergh, 2020). 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). The operationalization of these paradigms and ensuring universal compliance across diverse sectors and jurisdictions constitutes a formidable challenge, with diverging stakeholder-specific and actor-oriented ethical perceptions (Jobin et al., 2019).

13.1. Data Protection Actors as AI Enforcement First-Movers: A Global Review

Generative AI, particularly Large Language Models (LLMs) such as ChatGPT, has sparked intense global regulatory scrutiny, leading to swift enforcement actions from data protection authorities concerned about privacy implications. Following ChatGPT's launch on November 30, 2022, the Italian Data Protection Authority (Garante) took significant regulatory action by temporarily banning ChatGPT in March 2023, citing potential violations of the European Union's General Data Protection Regulation (GDPR).[1] This decision marked one of the first instances of a regulatory body imposing such direct limitations on generative AI in response to data protection concerns. The Garante's decision was grounded in Article 58(2)(f) of the GDPR, which empowers authorities to impose temporary or definitive limitations on data processing where significant compliance concerns arise. The authority's investigation highlighted multiple GDPR compliance issues, including potential shortcomings in data accuracy, transparency, and age verification measures—elements critical for lawful data processing under the GDPR (Lomas, 2023). The ban was lifted in April 2023 after OpenAI implemented compliance measures aimed at improving user transparency, instituting age checks, and enhancing data-handling protocols.

The Italian action against ChatGPT resonated across Europe, prompting similar inquiries by other EU and non-EU data protection authorities. France's Commission Nationale de l'Informatique et des Libertés (CNIL), Spain's data protection agency (AEPD), and Japan's Personal Information Protection Commission each initiated investigations, reflecting an emerging global consensus on the need for regulatory scrutiny of generative AI (Parodi & Orusov, 2023). These investigations underscore the broader regulatory concern about whether LLMs and similar AI technologies can align with existing data protection frameworks, which mandate transparency, purpose limitation, and explicit consent for data processing. For instance, the critique points to a persistent legal challenge: until AI developers adopt robust data protection measures compatible with the GDPR and other data protection frameworks, LLMs may struggle to achieve full compliance with privacy and/or data protection systems (Belli, 2023).

[1] See Italian Data Protection Authority. *Temporary ban of ChatGPT due to privacy concerns*. March 2023. Available at: https://www.garanteprivacy.it/home/docweb/-/docweb-display/docweb/9870847. Accessed on: 16 Nov. 2024.

In a similar and global vein, Austria saw the advocacy group NOYB file a complaint with its Data Protection Authority in April 2024, alleging that ChatGPT had provided inaccurate information about a public figure and failed to correct it or disclose data sources, potentially violating EU privacy regulations (Chee, 2024). This complaint echoes broader European concerns about the accountability and reliability of AI systems, especially as they interact with sensitive personal data. To address these complex compliance issues, the European Data Protection Board (EDPB) has established a task force to coordinate investigations and enforcement actions concerning ChatGPT across EU member states (Goujard, 2023). Drawing from established principles of EU administrative law and regulatory harmonization, the European task force endeavored to cultivate inter-agency coordination for AI oversight in a pre-establishment context where OpenAI operated without the benefit of the GDPR's one-stop-shop mechanism through January 2024 (European Data Protection Board, 2024). This regulatory landscape shifted materially when OpenAI incorporated its Irish subsidiary in February 2024 (Digital Watch Observatory, 2024), thereby centralizing future supervisory authority in Ireland's Data Protection Commission, though this administrative reorganization left undisturbed the continuing investigations initiated by various national authorities prior to the establishment date.

In 2024, the Brazilian National Data Protection Authority (ANPD) also ordered Meta to cease using data from Brazilian users to train its AI models, citing potential privacy violations and setting a five-day compliance deadline with daily fines for noncompliance (Laier, 2024). The UK ICO has also issued specific guidance on AI and data protection, with an emphasis on fairness, transparency, and accountability in AI systems. This guidance includes requirements for conducting Data Protection Impact Assessments (DPIAs) and ensuring compliance with core data protection principles (ICO, 2023).

The regulatory trajectory of artificial intelligence in China exemplifies a paradigm of sector-specific yet emergent accountability, akin to that observed in data protection regimes. This trajectory unfolds through a progressive framework, beginning with ethics-based soft norms, advancing to sectoral hybrid governance mechanisms, and culminating in the specialized regulatory oversight of generative AI. The evolving conceptualization of accountability among diverse stakeholders is evident in key developments such as the Expert Draft Proposal of 2023, which was subsequently integrated into the (Chinese) Model Artificial Intelligence Law (CASS Research Group on AI Ethics and Governance, 2023). This Model Law saw a revised version released in early 2024 (Zhou et al., 2024), alongside the publication of another Scholarly Draft Proposal of the same year (Zhang et al., 2024). The divergence between these two proposals highlights ongoing debates and unresolved issues within the legislative process, underscoring the dynamic and iterative nature of AI governance in China.

13.2. The Hybrid Regulatory Approach

In the current landscape, where global frameworks for AI governance remain in an embryonic phase of discursive development, influential stakeholders have begun to articulate positions on regulatory oversight. Notably, the CEO of OpenAI expressed support for regulatory scrutiny concerning AI/algorithmic safety during U.S. congressional hearings (Kang, C., 2023). Such statements underscore aspirations for a unified global framework for

AI governance. However, achieving such consensus is likely a protracted process, creating opportunities for enterprises to engage in regulatory arbitrage (Pollman, E., 2019). Concurrently, while the goal is to uphold technological neutrality while mandating the effective implementation of technologies, existing techno-neutral principles—such as those enshrined in the EU's General Data Protection Regulation (GDPR)—are facing increasing challenges (Wong, J., & Henderson, T., 2019). This tension highlights the complexities of balancing neutrality with the imperative for actionable regulatory measures in the AI domain.

Currently, these laws generally function as an interim regulatory framework for artificial intelligence. However, this arrangement highlights significant gaps in the enforcement tools available to regulators, particularly in addressing issues such as disinformation, defamation, and intellectual property infringements arising from the use of generative AI. In contrast, over 160 jurisdictions worldwide have implemented data protection legislation.[2] Yet, when it comes to generative AI and its impact on the collection, use, and dissemination of personal data—exemplified by platforms like ChatGPT—substantial legal shortcomings persist (Burgess, M., 2023). These deficiencies underscore the pressing need for more robust and tailored regulatory instruments to address the unique challenges posed by generative AI technologies.

13.2.1. "Soft" Ethics

In the context of China's governance framework for artificial intelligence, the initial regulatory approach aligned closely with prevailing international trends, emphasizing ethics as a form of soft law operating in parallel with progressive regulatory initiatives. To illustrate, in 2019, the Ministry of Science and Technology established the Next Generation Artificial Intelligence Governance Committee (Zeng, Y., 2020). On June 17 of the same year, the Committee issued the Principles for the Governance of New Generation Artificial Intelligence – Developing Responsible Artificial Intelligence (Next Generation Artificial Intelligence Governance Committee, 2019).

Building upon these foundations, the Guidelines on Strengthening the Governance of Technological Ethics, proposed in 2021 and officially launched in 2022, articulated key ethical principles. These include enhancing human well-being, respecting the sanctity of life, upholding fairness and justice, exercising prudent risk management, and maintaining openness and transparency. Additionally, the guidelines delineate specific responsibilities for innovative actors, including researchers, in advancing the governance of technological ethics (General Office of the CCP Central Committee & State Council Office, 2022). This multi-tiered framework underscored the then centrality of ethical considerations in China's evolving AI governance paradigm.

[2] 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/.

Furthermore, the 2021 Ethical Standards for the New Generation of Artificial Intelligence explicitly establishes foundational ethical norms for artificial intelligence. Simultaneously, it proposes a comprehensive array of management standards, research and development benchmarks, supply chain norms, and usage guidelines for AI applications. These standards aim to integrate ethical principles throughout the entire lifecycle of AI systems, from inception to deployment and use, thereby ensuring a consistent ethical framework governs all stages of AI development and utilization (Next Generation Artificial Intelligence Governance Committee, 2021).

However, the development of legislative frameworks for technology ethics in China has accelerated, incorporating ethical standards as procedural obligations within the AI ethics review system. For instance, the 2023 Provisional Measures for Science and Technology Ethics Review standardizes basic procedures, criteria, and conditions for ethics review, marking a significant evolution in China's regulatory system for AI ethics governance (Ministry of Science and Technology et al., 2023). These measures address ethical review and oversight for a wide range of scientific activities, including:

1. Research involving human participants, human biological samples, or personal data.
2. Activities not directly involving humans or animals but posing potential risks to health, the environment, public order, or sustainability.

The scope thus broadly encompasses nearly all technological activities, including those related to AI. Under the procedural requirements, entities engaging in relevant scientific activities must:

1. Establish or commission an ethics review committee for oversight.
2. Submit certain high-impact activities—such as the development of algorithms with public opinion mobilization or societal guidance capabilities, or autonomous decision-making systems in safety-critical contexts—for expert review by local or sectoral authorities.

Consequently, ethics review now operates as a prerequisite alongside algorithm filing and safety assessments for launching large AI models or services. This framework underscores China's robust approach to embedding ethical oversight in the development and deployment of advanced technologies.

13.2.2. Content Security

China's initial regulatory framework for algorithmic governance emerged through a hierarchical sequence of policy instruments, beginning with broad Party declarations and culminating in specific administrative measures. The foundational document in this regulatory cascade was the Chinese Communist Party (CCP) Central Committee's "Outline for Establishing a Rule-of-Law-Based Society (2020–2025)" (Central Committee of the Chinese Communist Party [CCP Central Committee], 2020). While this Party document lacked direct legal enforceability—consistent with China's distinction between Party

directives (党内法规) and state law (国家法律)—it nevertheless served as a crucial policy signal that shaped subsequent regulatory development. Notably, the Outline specifically identified algorithmic recommendations and synthetic media as emerging domains requiring legal governance, marking the Party-state's first high-level recognition of these technologies as subjects for regulatory intervention (CCP Central Committee, 2020).

This Party-level policy crystallized into more concrete regulatory action through the "Guiding Opinions on Strengthening Comprehensive Governance of Internet Information Service Algorithms" (Cyberspace Administration of China et al., 2021). This document, while still technically non-binding, represented a significant operational advance as it was jointly issued by the Cyberspace Administration of China (CAC) and several other administrative bodies, demonstrating an inter-agency consensus on regulatory approach. The CAC's leading role in promulgating these Guiding Opinions reflected its institutional emergence as China's primary algorithmic governance regulator, a position later reinforced through subsequent binding regulations.

China's algorithmic governance framework crystallized into binding law through the "Provisions on the Management of Algorithm Recommendation for Internet Information Services" (CAC et al., 2021), a departmental regulation that marked the transition from aspirational guidance to enforceable rules. These Provisions established a comprehensive regulatory architecture encompassing five distinct categories of algorithmic systems: (1) generation and synthesis algorithms, (2) personalized recommendation algorithms, (3) sorting and selection algorithms, (4) retrieval and filtering algorithms, and (5) scheduling and decision-making algorithms. For each category, the Provisions mandated specific compliance obligations, including: internal governance mechanisms, algorithmic evaluation and verification protocols, explicit and implicit labelling,[3] public disclosure requirements, user autonomy safeguards, and mandatory registration in a central algorithmic registry.

Building upon this foundation, China extended its regulatory reach to address the specific challenges posed by synthetic media through the "Provisions on the Management of Deep Synthesis Internet Information Services" (CAC et al., 2022). These Provisions offered the first statutory definition of "deep synthesis technology" in Chinese law, characterizing it as the application of generative algorithms—specifically including deep learning and virtual reality technologies—to produce synthetic digital content across multiple modalities: text, images, audio, video, and virtual environments. This definition's breadth reflected a conscious regulatory choice to establish comprehensive oversight over the full spectrum of synthetic media technologies.

[3] It is noteworthy that the CAC has formulated more detailed rules with regard to labeling AI-generated content, see Cyberspace Administration of China (CAC). (2024). *Measures for Labeling of AI-Generated Synthetic Content (draft for solicitation of comments)*. Available in English (Unofficial Translation) at: https://www.chinalawtranslate.com/en/ai-content-labels/. Accessed on: November 16, 2024.

13.2.3. Data Protection

As briefed above, on one hand, Article 24 of the Personal Information Protection Law (PIPL),[4] 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:

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

[4] 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/.

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.

13.3. Accountability in the Chinese AI "Laws"

China made a landmark move toward AI accountability by introducing a draft regulation on Generative AI on April 11, 2023, subsequently formalizing it as the Interim Measures for the Administration of Generative Artificial Intelligence Services on July 13, 2023. This regulation represents a pivotal effort to harmonize technological innovation with robust compliance requirements. It adopts a sophisticated governance framework that addresses critical issues such as data protection, intellectual property rights, and content security. By embedding these priorities into the regulatory structure, the measures reflect a deliberate attempt to foster AI's growth while ensuring that its development and deployment align with ethical and legal standards. This balanced approach underscores China's top-down intent to navigate the complexities of AI regulation in a rapidly evolving technological landscape.

Simultaneously, the Model AI Law proposed by scholars from the Chinese Academy of Social Sciences (CASS) represents a foundational step in China's approach to AI governance, emphasizing both regulation and development. Initially published in August 2023, it has undergone iterative updates, with a minor revision (v1.1) in September 2023 and a more substantial update (v2) released in April 2024. The law envisions the establishment of a new authority, the China Administration of AI (CAAI), tasked with overseeing AI safety, ethics, and innovation. It introduces a tiered regulatory framework, with stringent licensing requirements for high-risk AI applications on a "negative list" and lighter oversight for others. Obligations for developers and providers include pre-deployment safety assessments, incident reporting, and bias prevention, with additional responsibilities for foundation model developers. To promote AI innovation, the law proposes special AI zones and tax incentives for safety investments. Notably, the draft aligns with international influences like the EU's AI Act while adapting to China's unique needs, including provisions for national security and global competition.

On the other hand, The draft Artificial Intelligence Law of the People's Republic of China (Scholarly Draft Proposal) emphasizes promoting innovation and ensuring safety in the

development, provision, and use of AI technologies. It outlines principles such as transparency, fairness, accountability, and human oversight, while mandating compliance with ethical and legal standards. The law introduces mechanisms for AI governance, including specialized infrastructure, data-sharing frameworks, and intellectual property protections. Critical AI systems face heightened regulatory requirements like annual risk assessments, traceability, and security disclosures. Additionally, the law supports talent cultivation, green development, and international cooperation while establishing both general and sector-specific liability frameworks and compliance incentives to mitigate risks.

13.4. 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), 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) 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), 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.[5]

13.5. The Case of the Chinese Model AI Law (from v.1 to v.2)

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".[6] Specifically, the accountability principle in Article 7 of the Model Law encompasses three distinct aspects of normativity.

[5] See Art.1, Interim Measures.

[6] See the English Translation of the Model Law (v.1.0) for reference at https://digichina.stanford.edu/work/translation-artificial-intelligence-law-model-law-v-1-0-expert-suggestion-draft-aug-2023/.

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 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 start-ups, 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 (v.1.0) 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 (v.1.0) 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). 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 Recital 27, EU AI Act.[7] 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.

[7] 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 Model AI Law v.2.0, released on April 16, 2024, introduces significant updates to refine AI governance, promote innovation, and reduce regulatory burdens. A key improvement is the differentiation between licensing requirements for AI activities on a "negative list" and a simpler registry process for other AI applications. This distinction aims to avoid excessive compliance demands while maintaining regulatory oversight. The updated version emphasizes, in particular, support for open-source AI development. It includes measures to promote open-source communities, such as specialized compliance guidance, reduced liability for contributors, and clear responsibility rules.[8] Article 22 (v.2.0) introduces tax incentives to encourage investments in safety-related R&D and equipment, providing a minimum 30% tax credit, with specific preferences for open-source safety governance to be determined by the State Council. Provisions on intellectual property and data governance are also enhanced. These include rules for using training data and personal information in AI research, as well as new protections for AI-generated products.

Meanwhile, safety and security obligations have been streamlined. The removal of Article 39 eliminates mandatory safety evaluations, reducing direct compliance burdens. However, safety is still addressed under other provisions like Article 34. Additionally, Article 35 focuses on security vulnerability management, encouraging third parties to report vulnerabilities and obliging developers to address, disclose, and guide users on such risks.

13.6. Conclusion

The chapter opens with a critical examination of the historical evolution of AI accountability, a concept that has long defied precise articulation due to the interplay of complex notions such as explainability, accuracy, reliability, and resilience. It underscores the inherent challenges posed by the autonomy and complexity of AI systems, which necessitate a meticulous and unambiguous delineation of responsibility for their decision-making processes. These challenges are further exacerbated by the rapid pace of technological advancements, which make the formulation of effective regulatory and ethical frameworks increasingly difficult in an environment of unprecedented AI innovation.

China’s transition to a hybrid regulatory framework emerges as a paradigmatic case study, highlighting the need for a cohesive and adaptive global governance structure for AI. At the same time, it underscores the difficulties in maintaining technological neutrality within existing legal frameworks. China’s contemporary approach integrates content security, data protection, and sector-specific regulations, particularly in the realm of generative AI. This hybrid model ostensibly balances the dual imperatives of promoting innovation while ensuring compliance, thereby serving as a strategic blueprint for aligning growth with governance.

[8] Like the EU AI Act, the updated v.2.0 incorporates best practices such as liability exemptions for open-source AI contributors (Article 71). These provisions aim to foster a collaborative and responsible AI development ecosystem.

The Model AI Law and its iterations exemplify this balance by enshrining the principle of accountability, systematically stratifying responsibilities across the AI value chain. It delineates specific obligations concerning responsibility allocation, data retention, transparency in disclosures, and mutual cooperation among stakeholders. These measures aim to foster a transparent, ethical, and accountable AI ecosystem, thereby harmonizing the demands of technological advancement with the imperatives of regulatory oversight.

By addressing the technical, ethical, and regulatory dimensions of AI accountability, the paper underscores the importance of proactive governance mechanisms, tailored obligations, and accountability principles that are embedded in technological design. China's evolving regulatory strategies are contextualized within this broader narrative, illustrating how jurisdictions are grappling with the dual challenges of fostering innovation and maintaining robust accountability frameworks.

As AI continues to transform industries and societies, the paper emphasizes the urgent need for operationalizable accountability mechanisms. These mechanisms must strike a delicate balance between regulatory adaptability and predictability to ensure effective governance in an era defined by the dynamic and disruptive capabilities of AI. This nuanced equilibrium is positioned as an essential cornerstone for fostering trust, mitigating risks, and ensuring the ethical deployment of AI technologies in an ever-evolving landscape.

References

Bambauer, J., & Zarsky, T. (2018). The Algorithm Game. *Notre Dame Law Review*, 94(1), 1–48.

Belli, L. (2023). Why ChatGPT does not comply with the Brazilian Data Protection Law. *MediaNama*. Available at: https://www.medianama.com/2023/05/223-chatgpt-brazilian-data-protection-law-ai-regulation/. Accessed on: July 29, 2024.

Bordt, S., Finck, M., Raidl, E., & von Luxburg, U. (2022). Post-Hoc Explanations Fail to Achieve their Purpose in Adversarial Contexts. In *2022 ACM Conference on Fairness, Accountability, and Transparency* (pp. 891–905). DOI: https://doi.org/10.1145/3531146.3533153.

Bradford, A. (2023). The Race to Regulate Artificial Intelligence. *Foreign Affairs*. Available at: https://www.foreignaffairs.com/united-states/race-regulate-artificial-intelligence. Accessed on: July 29, 2024.

Burgess, M. (2023). ChatGPT Has a Big Privacy Problem. *Wired*. Available at: https://www.wired.com/story/italy-ban-chatgpt-privacy-gdpr/. Accessed on: July 29, 2024.

CAC, MIIT, MPS, & SAMR. (2021). *Provisions on the Management of Algorithm Recommendation for Internet Information Services*. Available at: https://www.gov.cn/zhengce/zhengceku/2022-01/04/content_5666429.htm. Accessed on: July 29, 2024.

CAC, MIIT, & MPS. (2022). *Provisions on the Management of Deep Synthesis Internet Information Services*. Available at: https://www.gov.cn/zhengce/zhengceku/2022-12/12/content_5731431.htm. Accessed on: July 29, 2024.

CAC, NDRC, ME, & MST. (2023). *Interim Measures for the Administration of Generative Artificial Intelligence Services*. Available at: http://www.cac.gov.cn/2023-07/13/c_1690898327029107.htm. Accessed on: July 29, 2024.

CAC. (2023). *Measures for the Administration of Generative Artificial Intelligence Services (Draft for Comment)*. Available at: http://www.news.cn/politics/2023-04/11/c_1129511663.htm. Accessed on: July 29, 2024.

CASS Research Group on AI Ethics and Governance. (2023). *Drafting Notes and Full Text of Artificial Intelligence Law (Model Law) 1.0 (Expert Proposal Draft)*. Available at: https://web.archive.org/web/20230822200944/https://redian.news/wxnews/536749. Accessed on: August 12, 2024.

Cavoukian, A. (2011). Privacy by Design: The 7 Foundational Principles. Available at: https://iab.org/wp-content/IAB-uploads/2011/03/fred_carter.pdf. Accessed on: July 29, 2024.

Central Committee of the Chinese Communist Party. (2020). *Outline for Establishing a Rule-of-Law-Based Society (2020–2025)*. Available at: https://www.gov.cn/zhengce/2020-12/07/content_5567791.htm. Accessed on: July 29, 2024.

Chee, F. Y. (2024). OpenAI's ChatGPT Targeted in Austrian Privacy Complaint. *Reuters*. Available at: https://www.reuters.com/technology/openais-chatgpt-targeted-austrian-privacy-complaint-2024-04-29/. Accessed on: November 16, 2024.

Coeckelbergh, M. (2020). Artificial Intelligence, Responsibility Attribution, and a Relational Justification of Explainability. *Science and Engineering Ethics*, 26(4), 2051–2068. DOI: https://doi.org/10.1007/s11948-019-00146-8.

Cyberspace Administration of China et al. (2021). *Guiding Opinions on Strengthening Comprehensive Governance of Internet Information Service Algorithms*. Available at: http://www.cac.gov.cn/2021-09/29/c_1634507915623047.htm. Accessed on: July 29, 2024.

Digital Watch Observatory. (2024). OpenAI Plans to Shift European Data Control to Ireland for GDPR Compliance. *Digital Watch Observatory*. Available at: https://dig.watch/updates/openai-plans-to-shift-european-data-control-to-ireland-for-gdpr-compliance. Accessed on: November 16, 2024.

Du, M., Liu, N., & Hu, X. (2019). Techniques for interpretable machine learning. *Communications of the ACM*, 63(1), 68–77. DOI: https://doi.org/10.1145/3359786.

English Translation of the Model Law. Available at: https://digichina.stanford.edu/work/translation-artificial-intelligence-law-model-law-v-1-0-expert-suggestion-draft-aug-2023/. Accessed on: July 29, 2024.

English Translation of the PIPL. Available at: https://digichina.stanford.edu/work/translation-personal-information-protection-law-of-the-peoples-republic-of-china-effective-nov-1-2021/. Accessed on: July 29, 2024.

European Data Protection Board. (2024). Report of the Work Undertaken by the ChatGPT Taskforce. Available at: https://www.edpb.europa.eu/system/files/2024-05/edpb_20240523_report_chatgpt_taskforce_en.pdf. Accessed on: November 16, 2024.

European Parliament's Version, in particular, Amendment 213. Available at: https://www.europarl.europa.eu/doceo/document/TA-9-2023-0236_EN.html. Accessed on: July 29, 2024.

General Office of the CCP Central Committee; State Council Office. (2022). *Guidelines on Strengthening the Governance of Technological Ethics*. Available at: https://www.gov.cn/zhengce/202212/content_6688372.htm. Accessed on: July 29, 2024.

Goujard, C. (2023). European Data Regulators Set up ChatGPT Task Force. *POLITICO*. Available at: https://www.politico.eu/article/european-data-regulators-set-up-chatgpt-taskforce/. Accessed on: November 16, 2024.

Grady, P. (2022). The EU Should Clarify the Distinction Between Explainability and Interpretability in the AI Act. *Center for Data Innovation*. Available at: https://datainnovation.org/2022/08/the-eu-should-clarify-the-distinction-between-explainability-and-interpretability-in-the-ai-act/. Accessed on: July 29, 2024.

ICO. (2023). Guidance on AI and Data Protection. Available at: https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/artificial-intelligence/guidance-on-ai-and-data-protection/. Accessed on: November 16, 2024.

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, 1(9). DOI: https://doi.org/10.1038/s42256-019-0088-2.

Kang, C. (2023). OpenAI's Sam Altman Urges A.I. Regulation in Senate Hearing. *The New York Times*. Available at: https://www.nytimes.com/2023/05/16/technology/openai-altman-artificial-intelligence-regulation.html. Accessed on: July 29, 2024.

Laier, P. (2024). Brazil Authority Suspends Meta's AI Privacy Policy, Seeks Adjustment. *Reuters*. Available at: https://www.reuters.com/technology/artificial-intelligence/brazil-authority-suspends-metas-ai-privacy-policy-seeks-adjustment-2024-07-02/. Accessed on: July 19, 2024.

Lomas, N. (2023). ChatGPT resumes service in Italy after adding privacy disclosures and controls. *TechCrunch*. Available at: https://techcrunch.com/2023/04/28/chatgpt-resumes-in-italy/. Accessed on: July 29, 2024.

Ministry of Science and Technology et al. (2023). *Measures for the Review of Science and Technology Ethics (Trial).* Available at: https://www.most.gov.cn/xxgk/xinxifenlei/fdzdgknr/fgzc/gfxwj/gfxwj2023/202310/t20231008_188309.html. Accessed on: June 1, 2024.

Next Generation Artificial Intelligence Governance Committee. (2019). *Principles for Next Generation AI Governance—Developing Responsible AI*. Available at: https://www.most.gov.cn/kjbgz/201906/t20190617_147107.html. Accessed on: July 29, 2024.

Next Generation Artificial Intelligence Governance Committee. (2021). *Ethical Standards for the New Generation of Artificial Intelligence*. Available at: https://www.most.gov.cn/kjbgz/202109/t20210926_177063.html. Accessed on: July 29, 2024.

Parodi, A., & Orusov, A. (2023). Governments race to regulate AI tools. *Reuters*. Available at: https://www.reuters.com/technology/governments-race-regulate-ai-tools-2023-08-22/. Accessed on: July 29, 2024.

Pollman, E. (2019). Tech, Regulatory Arbitrage, and Limits. *European Business Organization Law Review*, 20(3), 567–590. DOI: https://doi.org/10.1007/s40804-019-00155-x.

Privacy Law & Business. Available at: https://www.privacylaws.com/reports-gateway/articles/int181/int181_2023/. Accessed on: July 29, 2024.

The Garante's decision. Available at: https://www.garanteprivacy.it/web/guest/home/docweb/-/docweb-display/docweb/9870847. Accessed on: July 29, 2024.

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. DOI: https://doi.org/10.1093/idpl/ipz008.

Zeng, Y. (2020). Promoting the ethics and governance of the new generation of artificial intelligence. *ST Daily*. Available at: http://www.stdaily.com/index/kejixinwen/2020-06/04/content_952415.shtml. Accessed on: July 29, 2024.

Zhang, L., et al. (2024). *Artificial Intelligence Law of People's Republic of China (Scholarly Proposed Draft)*. Available at: https://bit.ly/3WKEBxn. Accessed on: August 12, 2024.

Zhou, H., et al. *The Model Artificial Intelligence Law (MAIL) v.2.0 - Multilingual Version*. DOI: https://doi.org/10.5281/zenodo.10974163.

Pre-print version of Belli L. Gaspar. W.B. The Quest for AI Sovereignty, Transparency and Accountability. Springer-Nature. (2025).

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-facing projects.

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[1]. In the process, the principle of transparency 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 (NITI Aayog., 2018). In a subsequent publication, the NITI Aayog laid out its recommended Principles for Responsible AI (NITI Aayog., 2021). 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 (NITI Aayog., 2021).

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

[1] 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.

governance. The Indian government has maintained a similar stand in other policy forums. For instance, 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 (The Hindu, 2023).

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 (NITI Aayog., 2022). As a part of this exercise, it undertook a deeper dive into one specific 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 was preceded by other types of policy documentation about the project. Notably, an explanatory note issued by the Ministry of Civil Aviation (MoCA) in 2018 (Ministry of Civil Aviation, Government of India, 2018) 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 (Jain A., 2022), (Kodali S., 2023) 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 14.1 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 14.2 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 14.3 summarizes the main findings and conclusions.

14.1. 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 (OECD, 2019) club transparency with explainability so as 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 (UNESCO, 2022) 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 (NITI Aayog, 2021). 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 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 (UNESCO, 2022).

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 lays out the expectations from different groups of stakeholders, such as the government, the private sector and research institutions (NITI Aayog, 2021). The NITI Aayog also places significant emphasis on the need for transparency in the procurement processes followed by the government while selecting a technology vendor and in sharing information about the error rates encountered during the project's implementation (NITI Aayog, 2022)

An annex to NITI Aayog's responsible AI principles contains examples of model transparency mechanisms (NITI Aayog, 2021). The list includes Google's Model Card Toolkit, Microsoft's datasheets for datasets, IBM's Fact Sheet project. All 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 (Tonfi Y., Masayuk O., Hiraku M., Kirihito Y. and Yuta N., 2023). However, as noted above, the various AI principles also 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 the actors involved in all stages from its design to implementation to evaluation (NITI Aayog, 2021). 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.

14.2. 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 (OECD, n.d.). 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 or backed by government agencies 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 DigiYatra project.

In August 2018 India's MoCA announced its plans to launch a biometric system for airport entry and boarding management procedures under the name of DigiYatra, which translates to mean 'digital journey' in Hindi (Ministry of Civil Aviation, 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 DigiYatra Central Ecosystem (NITI Aayog, 2022). The shareholding of this entity was held by the Airport Authority of India, a statutory authority under the MoCA and six other companies that operate as special purpose vehicles (SPVs) for the operation of airports in the cities of Cochin, Bengaluru, Delhi, Hyderabad and Mumbai (Digi Yatra Foundation, n.d.). The project was launched at three airports, New Delhi, Varanasi, and Bengaluru, in December 2022 and has since then been extended to several other cities (Ministry of Civil Aviation, Government of India., 2023).

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 (AWS) 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 (Ministry of Civil Aviation, Government of India., 2023) While the selection of Dataevolve Solution took place through an open challenge, the mode and specifics relating to the involvement of AWS remains less clear. AWS is reported to be partnering with Dataevolve for the execution of the project (Money Control., 2023). This includes reliance on tools like Amazon Cognito to authenticate passengers with an access token and AWS Lambda for password generation and verification (Amazon Web Services, 2023). In addition, each airport authority is selecting a technology implementation partner for giving effect to the facial recognition system on the ground (The Print., 2023). 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 think tank that claims to have 'been intimately involved in' the Digi Yatra project (Singh, S., 2022). 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 (NITI Aayog, 2022).

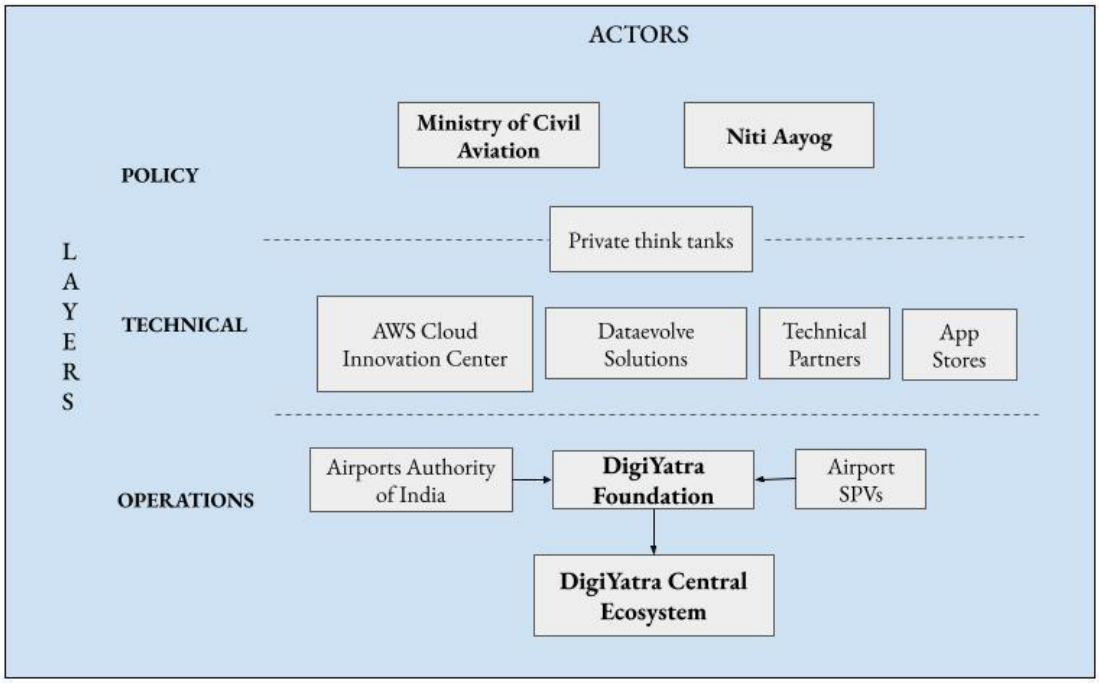


Figure 1: Actor map of India's Digi Yatra System

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 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. One of the information points that has come out recently is that DigiYatra has reduced 'processing times from an average of 15 seconds to approximately 5 seconds at airport entry gates' (Amazon Web Services, 2023). Although this is being positioned as an achievement, a more deliberative discussion based on the results of pilot studies may have highlighted different perspectives. Notably, this would include perspectives that do not regard the saving of 10 seconds as a proportionate response to such mass deployment of biometric technology.

In terms of technical design, the Digi Yatra Foundation reports that the solution is built on World Wide Web Consortium (W3C) standards using Self Sovereign Identity, Verifiable Credentials, Decentralized Identifiers and a Distributed Ledger for building trust between the ecosystem participants (Digi Yatra Foundation, n.d.). As for its technical

implementation, 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 (DigiYatra Foundation E-Procurement, n.d.). However, as discussed above, the technical partners involved in the project included a number of other entities, like AWS 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 centred 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.

Next, 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 DigiYatra is that since the DigiYatra 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 (Saravati N.T., 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.

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.

Finally, there is the critical aspect of technical and operational transparency related to the treatment of personally identifiable information under the DigiYatra system. The government's stance on this has been that the system functions by having the passenger's personal information stored on a secured walled on their phone and not in any central repository (Ministry of Civil Aviation, 2023). Further, the data is transferred from the user's phone to the airport's system only at the time of travel and current policies provide that it will be removed from the airport's system after 24 hours from the flight departure. The government argues that due to these safeguards the system does not pose any privacy concerns.[2]

However, the system's privacy terms and the DigiYatra policy can be changed unilaterally at any point by the government or the DigiYatra Foundation. For instance, the original policy document put out by the MoCA in 2018 stated that the passenger's biometric data would be deleted from the airport's system within 1 hour of the flight (Ministry of Civil Aviation, 2018) but this has now been extended to 24 hours. Further, the 2018 policy also said that the data would continue to remain in the DigiYatra system (as opposed to the

[2] Ibid.

airport's systems) even after that period.[3] The fluctuating nature of the safeguards coupled with the lack of any public accountability measures to verify their adherence adds to the inherent risks associated with the use of sensitive personal data.

14.3. 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 DigiYatra 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 transparency was mainly in the form of putting out information about the project in the public domain and in the procurement of technical implementation partners. The paper then pointed to four main gaps in the discharge of transparency obligations across the system's layers.

First, it was observed that the policy transparency surrounding the DigiYatra 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 (Saravati N.T., 2023).

Fourth, the fluctuating nature of the project's privacy safeguards, which can easily be varied with an amended policy document, coupled with the absence of accountability measures to monitor actual adherence with the policies, poses cause for concern.

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 tends to remain on algorithmic or technical transparency while ignoring the procedural and administrative elements. This becomes particularly relevant for large, public-

[3] Ibid.

facing, AI systems that involve a complex set of actors, with differential transparency expectations from the system's participants.

References

Andrada G, Clowes R.W., Smart P.R. Varieties of transparency: exploring agency within AI systems. AI & Society, 38, p. 1321–1331, 2023. Available at: <https://link.springer.com/article/10.1007/s00146-021-01326-6>.

Digi Yatra Foundation. About us. Available at: <https://digiyatrafoundation.com/>.

Digi Yatra Foundation. Frequently asked questions. Available at: <https://digiyatrafoundation.com/faq>.

DigiYatra Foundation E-Procurement. Available at: <https://digiyatra.procure247.com/home>.

Jain A. The dangers of DigiYatra & facial recognition enabled paperless air travel. Internet Freedom Foundation, 18 jan. 2022. Available at: <https://internetfreedom.in/dangers-of-digiyatra/#:~:text=The%20scheme%20aims%20to%20make,any%20remedies%20or%20regulatory%20framework>.

Kodali S. How 'Digi Yatra' Can Potentially Be Used to Police Your Travel. The Wire, 13 fev. 2023. Available at: <https://thewire.in/rights/digi-yatra-privacy-biometric-travel>.

Ministry of Civil Aviation, Government of India. Digi Yatra Biometric Boarding System: Reimagining Air Travel. Version 5.2, 9 ago. 2018. Available at: <https://www.civilaviation.gov.in/sites/default/files/Digi%20Yatra%20Policy%2009%20Aug%2018.pdf>.

Ministry of Civil Aviation, Government of India. Digi Yatra to be launched at 6 more airports. Press Information Bureau, 11 ago. 2023a. Available at: <https://pib.gov.in/PressReleaseIframePage.aspx?PRID=1947913>.

Ministry of Civil Aviation, Government of India. Digi Yatra to be implemented at Kolkata, Pune, Vijayawada and Hyderabad Airports by March 2023. Press Information Bureau, 2 fev. 2023b. Available at: <https://pib.gov.in/PressReleaseIframePage.aspx?PRID=1895743>.

Ministry of Civil Aviation. Under Digi Yatra, passengers' data is stored in their own device and not in centralized storage. Press Information Bureau, 16 mar. 2023. Available at: <https://www.pib.gov.in/PressReleasePage.aspx?PRID=1907479>.

Money Control. This Hyderabad-based startup is behind the airport walk-in app Digi Yatra. 2 fev. 2023. Available at: <https://www.moneycontrol.com/news/business/this-hyderabad-based-startup-is-behind-the-airport-walk-in-app-digiyatra-9989271.html>.

NITI Aayog. Discussion Paper: National Strategy for Artificial Intelligence. 2018. Available at: <https://indiaai.gov.in/documents/pdf/NationalStrategy-for-AI-Discussion-Paper.pdf>.

NITI Aayog. Responsible AI Adopting the Framework: A Use Case Approach on Facial Recognition Technology. nov. 2022. Available at: <https://niti.gov.in/sites/default/files/2023-03/Responsible-AI-AIForAll-Approach-Document-for-India-Part-Principles-for-Responsible-AI.pdf>.

NITI Aayog. Responsible AI Approach Document for India: Part 1 - Principles for Responsible AI. 2021a. Available at: <https://www.niti.gov.in/sites/default/files/2021-02/Responsible-AI-22022021.pdf>.

NITI Aayog. Responsible AI Approach Document for India: Part 2 - Operationalizing Principles for Responsible AI. 2021b. Available at: <https://www.niti.gov.in/sites/default/files/2021-08/TowardsResponsibleAI-newReport3.pdf>.

OECD. Catalog of Tools & Metrics for Trustworthy AI. Available at: <https://oecd.ai/en/catalogue/tools>.

OECD. Recommendations of the Council on Artificial Intelligence. 21 maio 2019. Available at: <https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449>.

Saravati N.T. India's Civil Aviation Ministry Says Information on Digi Yatra Cannot Be Sought under Right to Information. Medianama, 3 mar. 2023. Available at: <https://www.medianama.com/2023/03/223-civil-aviation-ministry-digi-yatra-right-to-information/>.

Singh, S. iSPIRT Foundation's Response to Union Budget 2022. iSPIRT Blog, 1 fev. 2022. Available at: <https://pn.ispirt.in/response-to-union-budget-2022/>.

The Hindu. No plan to regulate AI: IT Ministry tells Parliament. 5 abr. 2023. Available at: <https://www.thehindu.com/news/national/no-plan-to-regulate-ai-it-ministry-tells-parliament/article66702044.ece>.

The Print. IDEMIA Selected as Technology Partner by DIAL for DigiYatra. 20 jun. 2023. Available at: <https://theprint.in/ani-press-releases/idemia-selected-as-technology-partner-by-dial-for-digiyatra/1634106/>.

Tonfi Y., Masayuk O., Hiraku M., Kirihito Y., Yuta N. AI Algorithm Transparency Toolkit: A Proposal for a Governance System to Enable Society to Accept and Benefit from AI-based Innovations. 2023.

UNESCO. United Nations Educational, Scientific and Cultural Organization Recommendation on the Ethics of Artificial Intelligence. 2022. Available at: <https://unesdoc.unesco.org/ark:/48223/pf0000381137>.

Pre-print version of Belli L. Gaspar. W.B. The Quest for AI Sovereignty, Transparency and Accountability. Springer-Nature. (2025).

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.

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 (Thomas, M., 2022). Therefore, making Artificial Intelligence trustworthy will contribute to making the Internet trustworthy.[1] For instance, as the Internet evolved, the face of Web 2.0 has been the intermediaries like social media platforms, search engines etc. (O'Neill, S., 2022) 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 (Rogers, J., 2023). 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.

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

[1] 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.

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.

15.1. 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 (NITI Aayog, 2018), OECD AI principles (OECD, 2019), G20 AI Principles (G20, 2019), Australia's AI Intelligence Ethics Framework and AI Ethics Principles (Australian Government, 2019), EU Ethics Guidelines for Trustworthy AI (European Commission, 2019), EU-US TTC Joint Roadmap for Trustworthy AI and Risk Management (European Commission, 2022), NIST's AI Risk Management Framework (National Institute of Standards and Technology, 2023), Germany, Artificial Intelligence Strategy 2018 (German Federal Government, 2020), Singapore National AI Strategy 2019 (Smart Nation Digital Government Office, 2019), USA's National Artificial Intelligence Research and Development Strategic Plan 2023 (National Science and Technology Council, 2023), France's AI for Humanity 2017 (Villani, C., 2018), European Union's Artificial Intelligence for Europe 2018 (European Commission, 2018), European Union's The Artificial Intelligence Act, 2023 (European Commission., 2021), United Kingdom's A Pro-Innovation Approach to AI Regulation 2023 (Department for Science, Innovation and Technology, 2023), Japan's Social Principles of Human-Centric AI 2019 (The Government of Japan, 2019), The Global Partnership on Artificial Intelligence's AI principles (The Global Partnership on Artificial Intelligence's AI principles, 2020), United Nations' Principles for Ethical Use of AI in UN 2022 (UN System Chief Executives Board for Coordination., 2022) , UNESCO Ethics of Artificial Intelligence (UNESCO, 2023), and other private sector frameworks (Schiff J, D., Borenstein, J., & Laas, K., 2021). 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 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 (Rogers, J., 2023). However, these fall through the cracks as we leave other players undiscussed. Therefore, in chapter three, we will discuss establishing an effective governance structure 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.

15.2. 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 (Maithon, R., 2023).

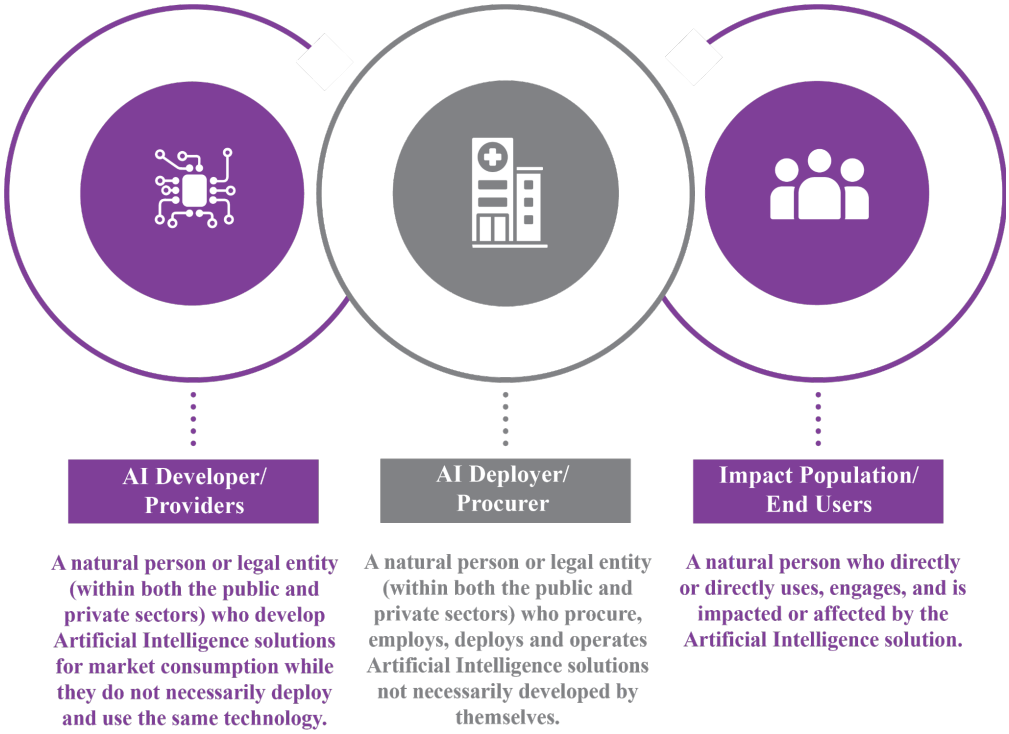
Therefore, through this chapter, we suggest a principle-based multi-stakeholder approach where we discuss various principles across the AI lifecycle bucketed and mapped to respective stakeholders within the AI ecosystem.[2]

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.[3]

Figure 1: Stakeholders

[2] 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.

[3] 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.



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 (Shankar, V., & Casovan, A., 2022), especially India (NITI Aayog, 2022).

15.3. 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[4] that can result from AI systems systematically and objectively (Metcalf J, Moss E, Watkins E, Singh R, and Elish M., 2021). 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.

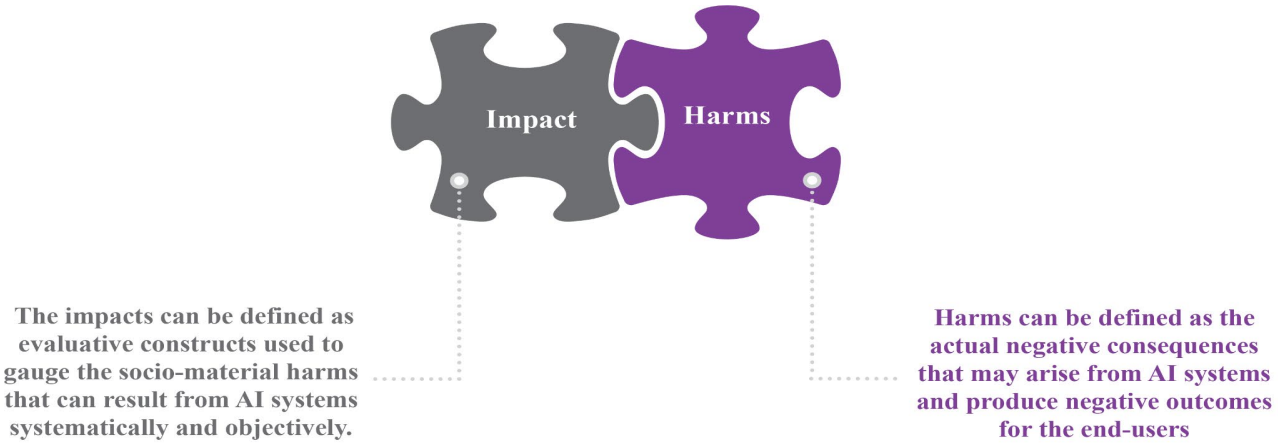
However, though the developed AI solutions are working as designed, adverse implications still crop out. This is where the other end of the puzzle, which is less discussed, comes into the picture, i.e., how AI deployers utilise the procured AI solutions for critical functions causing tangible and intangible harms (Horowitz, A., & Selbst, A., 2022). 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,

[4] 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.

real-life harms of such outcomes only translate into action when AI deployers blindly use the same for making real-life decisions.[5]

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 tackled appropriately (Ryan, M., 2020). 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.[6]

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 (European Commission., 2020). 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[7] "impact" and

[5]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.

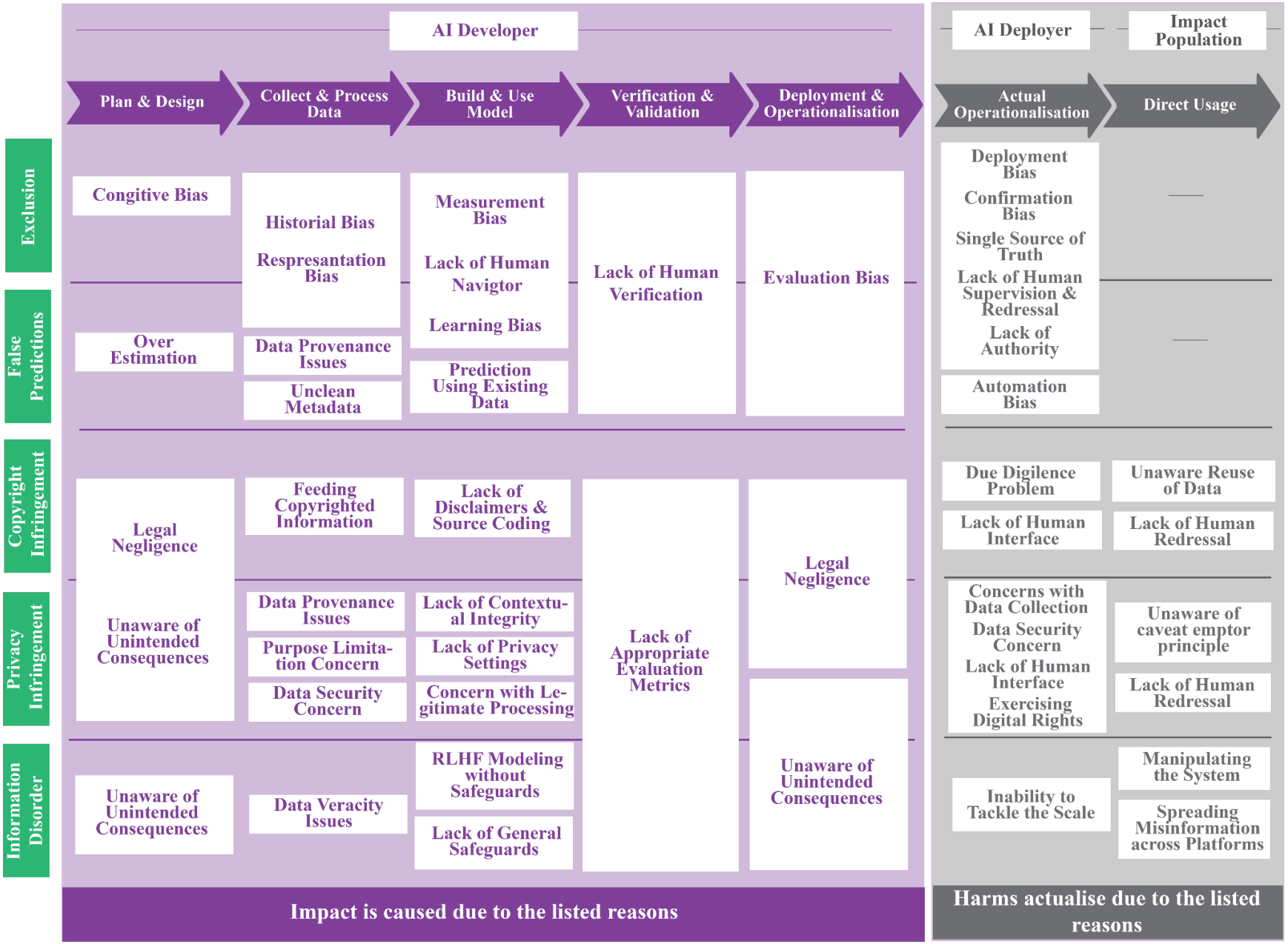
[6] Discussed in detail below

[7] Advised by OECD and NIST AI lifecycle, however, slightly improvised to fit the model suggested in this paper.

"harm" emerge 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.

Figure 3: Mapping Impact and Harms Across AI Lifecycle



15.3.1. 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 (Klein, A., 2022). Where individuals who attended colleges/universities dedicated to historically vulnerable populations have been quoted a higher interest rate and were denied credit (Klein, A., 2022).

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.

15.3.2. 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 patrols to the neighbourhoods where discriminated populations reside. The incorrect crime predictions also feed into the system, creating a vicious cycle (Sachoulidou, A., 2023). 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 (Goodman, R., 2023).

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.

15.3.3. Copyright Infringement

A problem that could have legal repercussions enforced through monetary claims is that of an AI system infringing intellectual property rights[1]. 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 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.

15.3.4. 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 (Sharma, D., 2023).

15.3.5. 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.

15.4. 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

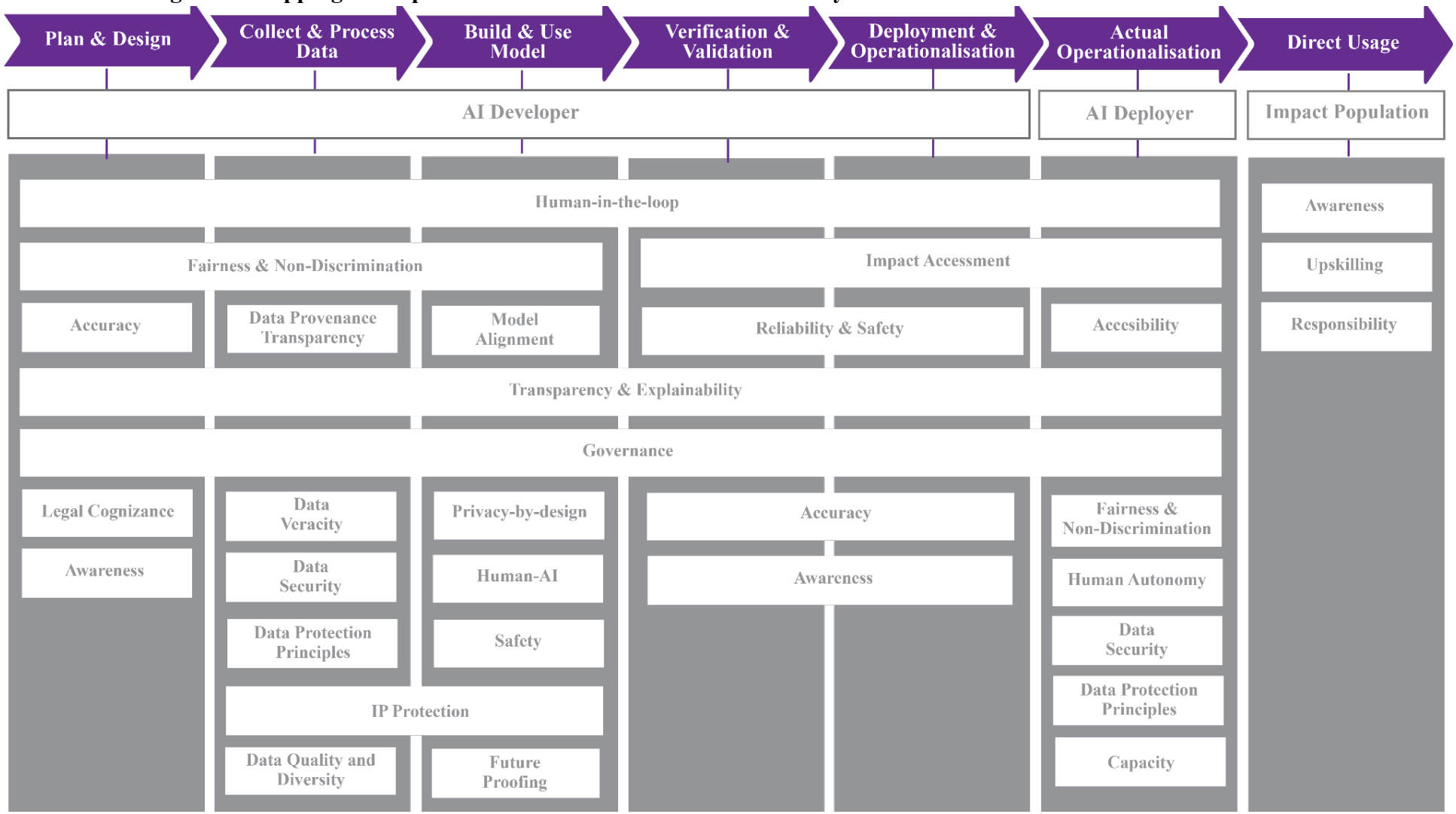
[1] For instance, Under the Indian IPR laws (Copyright Act 1957, Indian Patents Act, 1960 etc.) Patent and Copyright holders may sue AI developers for using their protected material for training foundation models. This has been observed across jurisdictions with prominent cases such as *Clakrson Law 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

the ecosystem level where various stakeholders follow different principles at different stages of the AI lifecycle.

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



15.5. 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.

15.5.1. 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.

15.5.2. 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.

15.5.3. 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.

15.5.4. 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.

15.5.5. 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.

15.5.6. 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.

15.5.7. 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.

15.5.8. 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.

15.6. 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.

15.6.1. 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.

15.6.2. 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 (OECD, 2021). Various jurisdictions have also emphasised this in the context of AI governance, where they believe concerted international-level regulatory cooperation is the way forward (Kerry, C. F., Meltzer, J. P., Renda, A., Engler, A., & Fanni, R., 2022).

15.6.3. 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 (UNESCO, 2021) 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 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.

15.7. 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.

References

Australian Government. Australia's AI Ethics Principles. 2019. Available at: <https://www.industry.gov.au/publications/australias-artificial-intelligence-ethics-framework/australias-ai-ethics-principles>.

Clarkson Law Firm v Open AI Case 3:23-cv-03199 in the United States of America. Available at: <https://clarksonlawfirm.com/wp-content/uploads/2023/06/0001.-2023.06.28-OpenAI-Complaint.pdf>.

Department for Science, Innovation and Technology. A pro-innovation approach to AI regulation. 2023. Available at:

<https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1146542/a_pro-innovation_approach_to_AI_regulation.pdf>.

European Commission. Artificial Intelligence for Europe. 2018. Available at: <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=COM:2018:237:FIN>.

European Commission. Ethics Guidelines for Trustworthy AI. High-Level Expert Group on Artificial Intelligence, 2019. Available at: <https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=60419>.

European Commission. The Artificial Intelligence Act. 2021. Available at: <https://artificialintelligenceact.eu/the-act/>.

European Commission. The ethics of artificial intelligence: Issues and initiatives. European Parliament, 2020. Available at: <https://www.europarl.europa.eu/RegData/etudes/STUD/2020/634452/EPRS_STU(2020)634452_EN.pdf>.

European Commission. TTC Joint Roadmap on Evaluation and Measurement Tools for Trustworthy AI and Risk Management. 2022. Available at: <https://ec.europa.eu/newsroom/dae/redirection/document/92123>.

G20. G20 AI Principles. 2019. Available at: <https://www.mofa.go.jp/policy/economy/g20_summit/osaka19/pdf/documents/en/annex_08.pdf>.

German Federal Government. National AI Strategy. 2020. Available at: <https://www.ki-strategie-deutschland.de/files/downloads/Fortschreibung_KI-Strategie_engl.pdf>.

Goodman, R. Why Amazon's automated hiring tool discriminated against women. American Civil Liberties Union, 2023. Available at: <https://www.aclu.org/news/womens-rights/why-amazons-automated-hiring-tool-discriminated-against>.

Horowitz, A., & Selbst, A. The fallacy of AI functionality. ACM Digital Library, 2022. Available at: <https://dl.acm.org/doi/fullHtml/10.1145/3531146.3533158>.

Kerry, C. F., Meltzer, J. P., Renda, A., Engler, A., & Fanni, R. Strengthening international cooperation on AI. Brookings, 2022. Available at: <https://www.brookings.edu/research/strengthening-international-cooperation-on-ai/>.

Klein, A. Credit denial in the age of AI. Brookings, 2022. Available at: <https://www.brookings.edu/research/credit-denial-in-the-age-of-ai/>.

Klein, A. Reducing bias in AI-based financial services. Brookings, 2022. Available at: <https://www.brookings.edu/research/reducing-bias-in-ai-based-financial-services/>.

Maithon, R. India needs a principles-based approach to regulating AI. Bharat Times, 2023. Available at: <https://news.bharattimes.co.in/india-needs-a-principles-based-approach-to-regulating-ai/>.

Metcalf J, Moss E, Watkins E, Singh R, and Elish M. Algorithmic Impact Assessments and Accountability: The Co-construction of Impacts. ACM Digital Library, 2021. Available at: <https://dl.acm.org/doi/pdf/10.1145/3442188.3445935>.

National Institute of Standards and Technology. Artificial Intelligence Risk Management Framework. NIST Technical Series Publications, 2023. Available at: <https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf>.

National Science and Technology Council. The National Artificial Intelligence R&D Strategic Plan 2023 Update. The White House, 2023. Available at: <https://www.whitehouse.gov/wp-content/uploads/2023/05/National-Artificial-Intelligence-Research-and-Development-Strategic-Plan-2023-Update.pdf>.

NITI Aayog. National Strategy for Artificial Intelligence #AIforAll. 2018. Available at: <https://niti.gov.in/sites/default/files/2019-01/NationalStrategy-for-AI-Discussion-Paper.pdf>.

NITI Aayog. RESPONSIBLE AI #AIFORALL Adopting the Framework: A Use Case Approach on Facial Recognition Technology. 2022. Available at: <https://www.niti.gov.in/sites/default/files/2022-11/Ai_for_All_2022_02112022_0.pdf>.

OECD. Recommendation of the Council on Artificial Intelligence. OECD Legal Instruments, 2019. Available at: <https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449>.

OECD. Why does international regulatory cooperation matter and what is it? OECD iLibrary, 2021. Available at: <https://www.oecd-ilibrary.org/sites/62c39d12-en/index.html?itemId=/content/component/62c39d12-en>.

O'Neill, S. What's The Difference Between Web 1.0, Web 2.0, And Web 3.0? MarTech Alliance, 2022. Available at: <https://www.lxahub.com/stories/whats-the-difference-between-web-1.0-web-2.0-and-web-3.0>.

Rogers, J. Artificial intelligence risk & governance. AI & Analytics for Business, 2023. Available at: <https://aiab.wharton.upenn.edu/research/artificial-intelligence-risk-governance/>.

Ryan, M. Artificial intelligence ethics guidelines for developers and users: Clarifying their content and normative implications. Discover Journals, Books & Case Studies | Emerald Insight, 2020. Available at: <https://www.emerald.com/insight/content/doi/10.1108/JICES-12-2019-0138/full/html>.

Sachoulidou, A. Going beyond the “common suspects”: To be presumed innocent in the era of algorithms, big data and artificial intelligence - artificial intelligence and law. SpringerLink, 2023. Available at: <https://link.springer.com/article/10.1007/s10506-023-09347-w>.

Schiff J, D., Borenstein, J., & Laas, K. AI ethics in the public, private, and NGO sectors: A review of a global document collection. Montreal AI Ethics Institute, 2021. Available at: <https://montrealethics.ai/ai-ethics-in-the-public-private-and-ngo-sectors-a-review-of-a-global-document-collection/>.

Shankar, V., & Casovan, A. A framework to navigate the emerging regulatory landscape for AI. The OECD Artificial Intelligence Policy Observatory - OECD.AI, 2022. Available at: <https://oecd.ai/en/wonk/emerging-regulatory-landscape-ai>.

Sharma, D. Samsung restricts use of generative AI tools after employees leak sensitive data using ChatGPT. India Today, 2023. Available at: <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>.

Smart Nation Digital Government Office. National Artificial Intelligence Strategy. Smart Nation Singapore, 2019. Available at: <https://www.smartnation.gov.sg/files/publications/national-ai-strategy.pdf>.

The Global Partnership on Artificial Intelligence's AI principles. Global Partnership on Artificial Intelligence - GPAI, 2020. Available at: <https://gpai.ai/about/>.

The Government of Japan. Social Principles of Human-Centric AI. 2019. Available at: <https://www.cas.go.jp/jp/seisaku/jinkouchinou/pdf/humancentricai.pdf>.

Thomas, M. The future of AI: How artificial intelligence will change the world. Built-In, 2022. Available at: <https://builtin.com/artificial-intelligence/artificial-intelligence-future>.

UN System Chief Executives Board for Coordination. Principles for the Ethical Use of Artificial Intelligence in the United Nations System. United Nations - CEB, 2022. Available at: <https://unsceb.org/sites/default/files/2022-09/Principles%20for%20the%20Ethical%20Use%20of%20AI%20in%20the%20UN%20System_1.pdf>.

UNESCO. Recommendation on the ethics of artificial intelligence. 2021. Available at: <https://en.unesco.org/about-us/legal-affairs/recommendation-ethics-artificial-intelligence>.

UNESCO. UNESCO adopts first global standard on the ethics of artificial intelligence. 2023. Available at: <https://www.unesco.org/en/articles/unesco-adopts-first-global-standard-ethics-artificial-intelligence>.

Villani, C. For A Meaningful Artificial Intelligence: French Strategy. AI for humanity, 2018. Available at: <https://www.aiforhumanity.fr/pdfs/MissionVillani_Report_ENG-VF.pdf>.

Pre-print version of Belli L. Gaspar. W.B. The Quest for AI Sovereignty, Transparency and Accountability. Springer-Nature. (2025).

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.

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 (Sidney Fussell, 2017). 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 diverse cultural, socioeconomic, or geographic contexts. 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 (Daniel S. Justin B., Jason B., Kelly L., 2020).

There is no question that order and appropriately regulated growth is preferable to the chaos that dominated the previous era of disruption (Schiff, 2020). 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 (Jacob M., Natalie A., 2020). 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 new standards to govern the development and deployment of AI technologies[1].

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.

16.1. 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 (Harry S., 2014). 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 (Solon Barocas, Andrew D. Selbst, 2016). There can be a number of subtle avenues for 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 (*id.*, 716-722).

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[2]. 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 (Kate C. & Ryan C., 2016). 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 (James Z. Londa S., 2018). All of these challenges are particularly severe in cases where there are significant geographic or cultural

[1] *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 centred 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 (2008).

[2] *See, e.g.,* Ruha Benjamin (2019); Anupam Chander (2017); Sonia K. Katyal (2019); Sandra G. Mayson (2019); Ifeoma Ajunwa (2020); Safiya Umoja Noble (2018); Solon Barocas & Andrew D. Selbst (2016, p. 671); Vidushi Marda & Shivangi Narayan (2020).

gaps between where models are built or where data is sourced and where the systems themselves operate (Shreya S., Yoni H., Eric B., James A., Jimbo W., D. Sculley, 2017).

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? (Virginia Eubanks, 2018).

Across advanced democracies, however, the bulk of scholarship on this issue has focused on bias and discrimination, and problematic impacts of these technologies on traditionally marginalized communities in the domestic context of the authors who are examining the issue.[3] As the next section demonstrates, this focus has coloured regulatory priorities in the AI governance space.

16.2. 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 (African Union, n.d.; Unesco, 2022). However, relatively few governments have actually passed normative, hard law frameworks to govern this space (Levin, Downes, 2023). 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 (European Commission, 2021, hereinafter EU AI Act).

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 (*idem*).

In the United States the most high-profile attempt to impose uniform standards has been the Biden administration's Blueprint for an AI Bill of Rights (The White House, n.d.). 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 (NIST, n.d.). 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 (See, e.g., Exec. Order No. 13859, 84 FR 3967, 2019).

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

[3] This is not intended to overlook contributions from Majority World scholars to the current discourse, such as, for example, Abeba Birhane (2020).

Intelligence Principles, which were adopted in 2019 (OECD, n.d.). 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" (GPAI, n.d.). 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 (*idem*). 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 (GPAI, *Community*, n.d.).

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 (Erie & Streinz, 2021). China, however, has also rolled out a number of ground-breaking policy initiatives, particularly through the powerful Cyberspace Administration of China, which recently imposed broad new rules to govern recommendation algorithms (Musser, 2022). 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. ("Emotional Entanglement...", 2021).

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. (Carmel, Paul, 2022) 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 (Sheehan, 2022).

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 (TBS-Canada, 2020). 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 ("Singapore's Approach...", 2022).

These initiatives, however, are clustered in the minority world (See, e.g., Chukwubuikem, Ater, 2023). 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.

16.3. 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" (Nou, Nyarko, 2023) While the "Brussels Effect" is probably the best-known framing for how local standards become globally influential, these impacts are not limited to EU processes (Bendiek, Stuerzer, 2023)

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 (Miller, 2003). 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 (*ib.*). 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 (Karanicolas, 2021).

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. (Margaret Kwoka, Michael Karanicolas, 2022) 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[4]. Similar diffusion pressures have been observed related to a number of other constitutional rights (Benedikt Goderis & Mila Versteeg, 2014).

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, 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 (Charles R. Shipan & Craig Volden, 2008). Both of these mechanisms are likely to produce regulatory postures which fail to optimally serve the needs of locals (*id.*). For example, concerns over such pressures among the states was a factor underlying the adoption of the commerce clause in the U.S. Constitution (*id.* at 849).

16.4. 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

[4] *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

governance challenges, including with regards to technical standards, ethics, governance, and best practices related to AI (African Union, 2022). 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 (Abeba Birhane, 2020). 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 2021, draws heavily from the OECD principles (NCAI, 2023). 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 (EU AI Act, title II, art. 5). 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 (Freedom House, n.d.).

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 (Human Rights Watch, 2021). 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 (Arthur Gwagwa, Erika Kraemer-Mbula, Nagla Rizk, Isaac Rutenberg, Jeremy de Beer, 2020). The exploitative labour relationships which underlie the development and improvement of 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 (Carmel, Paul, 2022).

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 (Mark Coeckelbergh, 2021). AI development is also undergirded by extractive supply chains whose environmental impacts are likewise centred in poor countries (Danae Tapia & Paz Peña, 2020). The development of AI requires enormous amounts of labour to label datasets, curate and moderate harmful content, and train and input data, which is likewise typically drawn from the global poor(Kate Crawford, 2021).

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 labour 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 (Julien Jeandesboz, 2021).

16.5. 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 inclusive development of AI technologies and equitable access to their benefits (ITU, n.d.).

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 (ITU, *Membership*, n.d.). 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 (Chinmayi Arun, eds. 2019). 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 (*id.* at 7-8). 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 (*id.* at 9). This is not to gloss over the traditional and ongoing role of European and U.S.-based companies in the global spyware trade (Privacy International, 2016). 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" (*id.* at 12).

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 materials for export, as well as data for companies based in wealthier parts of the world to extract in order to improve their products (Paola Ricaurte, 2019). 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 (*id.*).

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[5].

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 database (Hash Sharing Consortium, n.d.). GIFCT was designed to foster collaboration between governments, the private sector, and civil society, though the latter has complained of a lack of transparency (Emma Llansó, 2019). 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 (Facebook, 2020). Though the Oversight Board is not technically a multistakeholder body, it has included significant engagement with civil society (Brent Harris, 2020).

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") (ICANN, 2019). 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[6]. ICANN's multi-stakeholder model includes spaces for engagement by governments through the Governmental Advisory Committee[7], engagement by civil society through the Non-Commercial Stakeholder Group[8], engagement by internet end users through the At-Large Advisory Committee (ICANN, *About us*, n.d.), and engagement by business interests through the Commercial Stakeholders Group (ICANN, *Commercial Stakeholder Group*, n.d.). 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

---

[5] 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).

[6] 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.

[7] See Governmental Advisory Committee, ICANN GOVERNMENTAL ADVISORY COMMITTEE (Nov. 8, 2019), https://gac.icann.org/, archived at https://perma.cc/Z4LT-3MCY.

[8] 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.

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 egalitarian 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 favour even in the context of a consensus-driven and multistakeholder process[9]. 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[10]. 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 (Chinmayi Arun, 2018). Any set of global, or even regional, standards, is bound to gloss over important contextual cues related to the specific cultural, linguistic, political, or social nature of AI's impact in a given place or time (Arthur Gwagwa, Erika Kraemer-Mbula, Nagla Rizk, Isaac Rutenberg, Jeremy de Beer, 2020). 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 (Abeba Birhane, 2020).

16.6. 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 (Tao Xu, Xiaoming Sheng, Tianyi Zhang, Huan Liu, Xiao Liang, & Ao Ding, 2018). Because this research drove the development of vehicles' safety features, it led to design choices which supported favourable crash survival outcomes among this demographic, at the cost of worse survival rates among

[9] 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).

[10] See Milton L. Mueller, Detaching Internet Governance from the State: Globalizing the IANA, 4 GEO. J. INT'L AFF. 35 (2014).

those with different body types[11]. Context matters, and a lack of appropriate context can have dire, even fatal consequences for those unlucky enough to be excluded from consideration.

It is a good thing that the world's most influential policymakers 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.

Policymakers, 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.

References

About GPAI. The Global Partnership on Artificial Intelligence. Available at: <https://gpai.ai/about/>.

African Union. AU data policy framework. February 2022. Available at: https://au.int/sites/default/files/documents/42078-doc-AU-DATA-POLICY-FRAMEWORK-ENG1.pdf.

African Union. The Digital Transformation Strategy For Africa (2020-2030). Available at: https://au.int/sites/default/files/documents/38507-doc-dts-english.pdf.

AI Risk Management Framework. National Institute of Standards and Technology. Available at: <https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf>.

Ajunwa, Ifeoma. The Paradox of Automation as Anti-Bias Intervention. Cardozo Law Review, 2020, vol. 41, p. 1671.

Alam, Shahidul. Majority World: Challenging the West's Rhetoric of Democracy. Amerasia Journal, 2008, vol. 34, p. 88.

Arun, Chinmayi. AI and the Global South: Designing for Other Worlds. In: Dubber, Markus D., Pasquale, Frank, & Das, Sunit (eds.). The Oxford Handbook of Ethics of AI. Oxford University Press, 2019.

Arun, Chinmayi. Rebalancing Regulation of Speech: Hyper-Local Content on Global Web-Based Platforms. Medium, 2018. Available at: <https://medium.com/berkman-klein-center/rebalancing-regulation-of-speech-hyper-local-content-on-global-web-based-platforms-1-386d65d86e32>.

[11] 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.

Avery, Daniel. Gay Dating App Grindr Still Leaking Users' Location Data, Report Indicates. Newsweek, 2019. Available at: <https://www.newsweek.com/grindr-location-leak-1453697>.

Barocas, Solon, & Selbst, Andrew D. Big Data's Disparate Impact. California Law Review, 2016, vol. 104, p. 671, 674.

Barocas, Solon, & Selbst, Andrew D. Big Data's Disparate Impact. California Law Review, 2016, vol. 104, p. 671.

Bendiek, Annegret, & Stuerzer, Isabella. 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. Digital Society, 2023, vol. 2, p. 5.

Benjamin, Ruha. Assessing Risk, Automating Racism: A health care algorithm reflects underlying racial bias in society. Science, 2019, vol. 366, p. 421.

Birhane, Abeba. Algorithmic Colonization of Africa. SCRIPTed, 2020, vol. 17, p. 389.

Blueprint for an AI Bill of Rights. The White House. Available at: <https://www.whitehouse.gov/ostp/ai-bill-of-rights/>.

Bremmer, Ian, & Suleyman, Mustafa. The AI Power Paradox Can States Learn to Govern Artificial Intelligence—Before It's Too Late?. Foreign Affairs, 2023, vol. 102, n. 5.

Carmel, Emma, & Paul, Regine. Peace and prosperity for the digital age? The colonial political economy of European AI governance. IEEE Technology and Society Magazine, 2022, vol. 41.

Chander, Anupam. The Racist Algorithm?. Michigan Law Review, 2017, vol. 115, p. 1023.

Coeckelbergh, Mark. AI for climate: freedom, justice, and other ethical and political challenges. AI and Ethics, 2021, vol. 1, p. 67. Dhar, Payal. The carbon impact of artificial intelligence. Nature Machine Intelligence, 2020, vol. 2, p. 423.

Commercial Stakeholder Group. ICANN. Available at: <https://gnso.icann.org/en/about/stakeholders-constituencies/csg>. Archived at: <https://perma.cc/CT4R-FM77>.

Crawford, Kate, & Calo, Ryan. There is a blind spot in AI research. Nature, 2016, vol. 538, p. 311.

Crawford, Kate. Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence. 2021, pp. 53-89.

Eke, Damian Okaibedi, Wakunuma, Kutoma, & Akintoye, Simisola. Responsible AI in Africa. 2023.

Emotional Entanglement: China's emotion recognition market and its implications for human rights. Article 19, 2021. Available at: <https://www.article19.org/wp-content/uploads/2021/01/ER-Tech-China-Report.pdf>.

Erie, Matthew S., & Streinz, Thomas. The Beijing Effect: China's 'Digital Silk Road' as Transnational Data Governance. N.Y.U. Journal of International Law & Politics, 2021, vol. 54, p. 1. Png, Marie-Therese. At the Tensions of South and North: Critical Roles of Global South Stakeholders in AI Governance. ACM Conference on Fairness, Accountability, and Transparency, 2022. Available at: <https://dl.acm.org/doi/10.1145/3531146.3533200>.

Eubanks, Virginia. Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor. 2018.

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. 2021.

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).

Freedom House. Freedom House's annual index of democratic health. Available at: <https://freedomhouse.org/countries/freedom-world/scores?sort=desc&order=Total%20Score%20and%20Status>.

Fussell, Sidney. Why Can't This Soap Dispenser Identify Dark Skin?. Gizmodo, 2017. Available at: <https://gizmodo.com/why-cant-this-soap-dispenser-identify-dark-skin-1797931773>.

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.

Goderis, Benedikt, & Versteeg, Mila. The Diffusion of Constitutional Rights. International Review of Law and Economics, 2014, vol. 39, p. 1.

Governmental Advisory Committee. ICANN Governmental Advisory Committee, 2019. Available at: <https://gac.icann.org/>. Archived at: <https://perma.cc/Z4LT-3MCY>.

GPAI. *About* GPAI, THE GLOBAL PARTNERSHIP ON GLOBAL INTELLIGENCE, https://gpai.ai/about/.

GPAI. *Community*, THE GLOBAL PARTNERSHIP ON GLOBAL INTELLIGENCE, https://gpai.ai/about/.

Gwagwa, Arthur, Kraemer-Mbula, Erika, Rizk, Nagla, Rutenberg, Isaac, & de Beer, Jeremy. Artificial intelligence (AI) deployments in Africa: Benefits, challenges and policy dimensions. African Journal of Information and Communication, 2020, vol. 26.

Harris, Brent. Preparing the Way Forward for Facebook's Oversight Board. Facebook, 2020. Available at: <https://about.fb.com/news/2020/01/facebooks-oversight-board>.

Hash Sharing Consortium. Global Internet Forum to Counter Terrorism. Available at: <https://gifct.org/joint-tech-innovation>.

Human Rights Watch. Uganda: Events of 2021. 2021. Available at: <https://www.hrw.org/world-report/2022/country-chapters/uganda>.

ICANN. About Us. ICANN AT-LARGE. Available at: <https://atlarge.icann.org/about/index>. Archived at: <https://perma.cc/3L22-XZ5P>.

ICANN. *Commercial Stakeholder Group*. Available at: https://gnso.icann.org/en/about/stakeholders-constituencies/csg, archived at https://perma.cc/CT4R-FM77.

Insperity OrgPlus 2012. ICANN, 2019. Available at: <https://www.icann.org/en/system/files/files/management-org-01may18-en.pdf>. Archived at: <https://perma.cc/4TRW-Z83N>.

International Telecommunication Union (ITU). Artificial Intelligence. Available at: <https://www.itu.int/en/action/ai/Pages/default.aspx>.

International Telecommunication Union (ITU). Membership. Available at: https://www.itu.int/hub/membership/how-we-are-governed/.

Jeandesboz, Julien. Technology, knowledge and the governing of migration. In: Carmel, E., Lenner, K., & Paul, R. (Eds.). Handbook on the Governance and Politics of Migration. 2021, p. 337.

Karanicolas, Michael. Even in a Pandemic, Sunlight Is the Best Disinfectant: COVID-19 and Global Freedom of Expression. Oregon Review of International Law, 2021, vol. 22, p. 101.

Karanicolas, Michael. The New Cybersquatters: The Evolution of Trademark Enforcement in the Domain Name Space. Fordham Intellectual Property, Media & Entertainment Law Journal, 2020, vol. 30, p. 399.

Katyal, Sonia K. Private Accountability in the Age of Artificial Intelligence. UCLA Law Review, 2019, vol. 66, p. 54.

Kwoka, Margaret, & Karanicolas, Michael. Overseeing Oversight. Connecticut Law Review, 2022, vol. 54, p. 657, 663.

Levin, Blair, & Downes, Larry. Who Is Going to Regulate AI?. Harvard Business Review, 2023. Available at: <https://hbr.org/2023/05/who-is-going-to-regulate-ai>.

Llansó, Emma. Platforms Want Centralized Censorship. That Should Scare You. Wired, 2019. Available at: <https://www.wired.com/story/platforms-centralized-censorship/>.

Marda, Vidushi, & Narayan, Shivangi. Data in New Delhi's Predictive Policing System. FAT '20: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, 2020. Available at: <https://doi.org/10.1145/3351095.3372865>.

Marda, Vidushi, & Narayan, Shivangi. Data in New Delhi's Predictive Policing System. Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, 2020. Available at: https://doi.org/10.1145/3351095.3372865.

Mayson, Sandra G. Bias In, Bias Out. Yale Law Journal, 2019, vol. 128, p. 2218.

Mchangama, Jacob, & Alkiviadou, Natalie. The Digital Berlin Wall: How Germany (Accidentally) Created a Prototype for Global Online Censorship - Act Two. Justitia, 2020. Available at: <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>.

Miller, Jonathan M. A Typology of Legal Transplants: Using Sociology, Legal History and Argentine Examples to Explain the Transplant Process. American Journal of Comparative Law, 2003, vol. 51, p. 839, 846.

Mozur, Paul. A Genocide Incited on Facebook, With Posts from Myanmar's Military. The New York Times, 2018. Available at: <https://www.nytimes.com/2018/10/15/technology/myanmar-facebook-genocide.html>. Archived at: <https://perma.cc/3Z2J-K6BA>.

Mueller, Milton L. Detaching Internet Governance from the State: Globalizing the IANA. Georgetown Journal of International Affairs, 2014, vol. 4, p. 35.

Musser, Micah. Don't Assume China's AI Regulations Are Just a Power Play. Lawfare, 2022. Available at: <https://www.lawfareblog.com/dont-assume-chinas-ai-regulations-are-just-power-play>.

NCAI. Egyptian Charter for Responsible AI. 2023. Available at: https://aicm.ai.gov.eg/en/Resources/EgyptianCharterForResponsibleAIEnglish-v1.0.pdf.

Nigeria: Events of 2021. Human Rights Watch. Available at: <https://www.hrw.org/world-report/2022/country-chapters/nigeria>.

NIST. *AI Risk Management Framework*. https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf.

Noble, Safiya Umoja. Algorithms of Oppression: How Search Engines Reinforce Racism. 2018.

Non-Commercial Stakeholder Group. ICANN, 2019. Available at: <https://gnso.icann.org/en/about/stakeholders-constituencies/ncsg>. Archived at: <https://perma.cc/KW4Y-45X5>.

Nou, Jennifer, & Nyarko, Julian. Regulatory Diffusion. Stanford Law Review, 2023, vol. 74, p. 897.

O'Flaherty, Kate. YouTube keeps deleting evidence of Syrian chemical weapon attacks. Wired, 2018. Available at: <https://www.wired.co.uk/article/chemical-weapons-in-syria-youtube-algorithm-delete-video>.

Obianyo, Chukwubuikem I., & Ater, Solomon Vendaga. A Critical Appraisal of the Legal Framework of Artificial Intelligence Governance in Nigeria. Journal of Private and Public Law, 2023, vol. 4, p. 48.

OECD AI Principles Overview. OECD Policy Observatory. Available at: <https://oecd.ai/en/ai-principles>.

Oversight Board Bylaws. Facebook, 2020. Available at: <https://about.fb.com/wp-content/uploads/2020/01/Bylawsv6.pdf>.

Privacy International. The global surveillance industry. July 2016. Available at: https://www.privacyinternational.org/sites/default/files/2017-12/global_surveillance_0.pdf.

Ricaurte, Paola. Data Epistemologies, Coloniality of Power, and Resistance. Television & New Media, 2019, vol. 20, p. 350, 358.

Schiff, Daniel, Biddle, Justin, Borenstein, Jason, & Laas, Kelly. What's Next for AI Ethics, Policy, and Governance? A Global Overview. AAAI/ACM Conference on AI, Ethics, and Society, 2020. Available at: <https://doi.org/10.1145/3375627.3375804>.

Shahidul Alam, *Majority World: Challenging the West's Rhetoric of Democracy*, 34 AMERASIA JOURNAL 88 (2008).

Shankar, Shreya, Halpern, Yoni, Breck, Eric, Atwood, James, Wilson, Jimbo, & Sculley, D. No Classification without Representation: Assessing Geodiversity Issues in Open Data Sets for the Developing World. NIPS 2017 Workshop on Machine Learning for the Developing World, 2017.

Sheehan, Matt. China's New AI Governance Initiatives Shouldn't Be Ignored. Carnegie Endowment for International Peace, 2022. Available at: <https://carnegieendowment.org/2022/01/04/china-s-new-ai-governance-initiatives-shouldn-t-be-ignored-pub-86127>.

Shipan, Charles R., & Volden, Craig. The Mechanisms of Policy Diffusion. American Journal of Political Science, 2008, vol. 52, pp. 840-848.

Singapore's Approach to AI Governance. Personal Data Protection Commission, 2022. Available at: <https://www.pdpc.gov.sg/Help-and-Resources/2020/01/Model-AI-Governance-Framework>.

Surden, Harry. Machine Learning and Law. Washington Law Review, 2014, vol. 89, pp. 87-90.

Tapia, Danae, & Peña, Paz. 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. Available at: <https://giswatch.org/node/6247>.

TBS-Canada. Directive on Automated Decision-Making. 2020. Available at: https://www.tbs-sct.canada.ca/pol/doc-eng.aspx?id=32592&section=html.

The White House. *Blueprint for an AI Bill of Rights*. Available at: https://www.whitehouse.gov/ostp/ai-bill-of-rights/.

UNESCO. Recommendation on the Ethics of Artificial Intelligence. UNESDOC Digital Library, 2022. Available at: https://unesdoc.unesco.org/ark:/48223/pf0000381137.

Vulnerability and Effectiveness of Occupant Protection Technologies for Older Occupants and Women. National Highway Traffic Safety Administration, 2013. Available at: <https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/811766>.

Welcome to ICANN!. ICANN, 2019. Available at: <https://www.icann.org/resources/pages/welcome-2012-02-25-en>. Archived at: <https://perma.cc/M8T9-2XCS>.

Xu, Tao, Sheng, Xiaoming, Zhang, Tianyi, Liu, Huan, Liang, Xiao, & Ding, Ao. Development and Validation of Dummies and Human Models Used in Crash Test. Applied Bionics and Biomechanics, 2018. Available at: <https://www.hindawi.com/journals/abb/2018/3832850/>.

Zou, James, & Schiebinger, Londa. Design AI so that it's fair. Nature, 2018, vol. 559.

Pre-print version of Belli L. Gaspar. W.B. The Quest for AI Sovereignty, Transparency and Accountability. Springer-Nature. (2025).

17. (Re)Examining the Concept of Regulation in AI Governance: Modest Efforts in Africa

Jake Okechukwu Effoduh

Introduction

The transformative power of artificial intelligence (AI) is undeniable, yet it carries with it a plethora of challenges that must be addressed to ensure its benefits are fully realized while mitigating inherent risks. In Kenya and Nigeria, where mobile money services are extensively utilized, AI-driven financial technology companies play a pivotal role in promoting much-needed financial inclusion. However, concerns have arisen regarding the privacy implications of these FinTechs' collection and utilization of personal data (Ifeanyi-Ajufo, Nnenna, 2022; Fundira, Edoun, Pradhan & Mbohwa, 2024)). In other countries like South Africa, discussions about AI applications in criminal justice, such as predictive policing algorithms, often intertwine with concerns about the absence of suitable regulatory frameworks to address algorithmic bias and discrimination against marginalized communities (Singh, Divya, June 2022). This duality positions AI as a subject of considerable regulatory interest. And just like how regulatory propositions were activated by other technologies in the past, (Picker, Colin B., 2001) there are concerns about how regulation should safeguard the public interest, maintain technology market integrity, and ensure that AI systems do not compromise shared extant norms.

The intense regulatory interest in AI is due to the technology's rapid development and the production of broad and generative applications across countries (some of which are unprecedented and can potentially cause irreversible impacts on people and society). For example, legal scholars have examined the implications of AI on public decision-making, raising concerns about social control and state power (Langford, Malcolm., 2020). Also, AI-based martial applications have initiated new multilateral interventions on how these AI technologies affect military cooperation (Hill, Steven., 2020). And then, with the challenges of applying traditional legal frameworks to novel technological scenarios, much is now considered about how the deployment of AI interacts with existing governance structures (Smith, Bryant Walker., 2020).

Therefore, the discourse around AI regulation (though relatively nascent) is already complex, involving many factors, including jurisdiction issues, governance, legal standards, and capacity to achieve international cooperation. While there may be some odd propositions around regulating AI (such as for the technology not to be regulated), ("Regulating AI is a mistake.", 2023), the dominant question is not whether AI should be regulated but rather how, by whom, and to what extent (Reed, Chris., 2018). And because AI is a deterritorial technology,[1] tensions exist as to whether regulation should be domestic, international, or both and if it should be top-down, bottom-up or both. Even the location of authority and capacity for AI regulation are still challenged across various jurisdictions. These are only, but a few of the many regulatory questions that require answers in the context of AI governance. This paper explores the imperative for such consideration by examining the justification of

[1] Globalization has made the application of AI "deterritorialized" and cutting across several jurisdictions. Therefore, state-centric approaches to governing the technology may be limiting.

regulation as a concept for AI governance, why it may be preferred, the type of regulation that may be most appropriate for AI, and some of its challenges.

17.1. Recognizing regulation as a concept in *luctatione et transformatione*[2]

Regulation is a highly contested term both in law literature and in society (Christel Koop & Martin Lodge, 2017). However, regulation is mainly conceived in several broad and abstract terms, characterized by interventions that can be intentional and direct, involving binding standard-setting, monitoring, and sanctioning, mainly by public-sector actors on private-sector activities (Black, Julia, and Dimity Kingsford Smith, 2002). As a concept, it has also majorly been a promulgation of authoritative rules with a mechanism for compliance, and as sustained, focused control by a public agency over activities valued by a state or community (Koop, Lodge, 2017). However, regulation is mainly characterized by state intervention in the private domain, aiming to improve citizen's ability to make choices without imposing restrictions on those choices (Barak Orbach, 2012).

Successful or otherwise, this has been a by-product of our imperfect constructed reality (*ibid.*). Many regulatory rules, for example, have been imposed on businesses to correct market failures, but some have been revealed as inefficiencies in bureaucratic decision-making, highlighting the contradictions and absurdities in state politics and public opinion (Joan Didion, 1979). Quite recently, however, regulation has evolved to include a transnational dimension where various actors participate in setting and enforcing rules. Several countries have since enforced various regulations by specialized agencies and over time, this has expanded from addressing single-sector issues to cross-sector problems like environmental protection, human rights, and consumer product safety (Marian Dohler, 2011).

In the last two decades, regulation has transformed from imposing rules to taking proactive measures (including deregulation to encourage competition). Therefore, regulation as a concept has been in flux – a complex and interdisciplinary field that lacks a shared understanding and is unconfined with disciplinary boundaries. It's an evolving field that is hugely challenged today by technological globalization, yet it still requires a consideration of various control mechanisms, including those that are indirect or unintentional (Koop, Lodge, 2017).

17.2. Choosing regulation above other governance efforts

Lawrence Lessig's work on the "Big Four" modalities of regulation provides a useful framework for understanding the multifaceted approach required to regulate AI (Lawrence Lessig, 1999). The Big Four represents the ways society can regulate behaviour, whether in cyberspace or the physical world. There seems to be no strict separation between these modes of regulation as combinations are possible, even as not doing anything is also a choice, the risks call for some type(s) of regulation to be considered. Originally, Lessig's framework was designed to explain internet regulation, but it also provides a valuable perspective for understanding AI regulation in Africa. Given the limited AI-specific legislation across the continent, it could be argued that AI in Africa is largely regulated as a general technology, similar to telecommunications or digital financial services. This broad approach contrasts with the European Union's more specialized, risk-based regulatory framework tailored

[2] "*In struggle and transformation*", in Latin.

specifically to AI systems. The EU AI Act, for instance, classifies AI applications by risk level (unacceptable, high, limited, and minimal) and applies obligations accordingly (European Commission, 2024). High-risk AI systems, under Articles 8-17, face the strictest requirements, including mandatory risk management, robust data governance, and extensive technical documentation (European Commission, 2024). Additionally, the Act includes provisions for general-purpose AI models with systemic risks, acknowledging AI's evolving nature and societal impacts.

In contrast, African nations are taking a more incremental approach to AI regulation. Some, like Egypt, incorporate AI into broader technology governance frameworks (Egypt Law Regulating and Developing the Use of Technology for Non-Banking Financial Activities, 2022), while others, such as Morocco, focus on data protection laws as indirect AI regulation (Morocco Law No. 09-08 on the Protection of Individuals regarding the Processing of Personal Data). Indeed, data privacy laws form the backbone of AI-related legislation in many African countries. Nigeria's Data Protection Act of 2023, though not specific to AI, imposes data protection principles that AI systems handling personal data must follow, addressing critical privacy and security issues (Nigeria Data Protection Act, 2023). Similarly, Kenya's Data Protection Act of 2019 provides regulatory oversight for AI systems processing personal information, highlighting data protection's growing role in African AI governance (Kenya Data Protection Act, 2019). These laws, often based on constitutional principles and regional frameworks like the SADC Model Law on Data Protection, mark an important step toward AI governance in Africa. While these frameworks offer initial safeguards, the effectiveness of this approach in meeting AI-specific regulatory needs remains a question, underscoring an evolving regulatory landscape across Africa.

Lessig argued that while the law is a powerful regulator, it's not the only force that shapes behaviour. Each modality – i.e. law, norms, market, and code offer a different mechanism for control and can be leveraged to achieve specific regulatory outcomes.

Lessig also emphasized that in cyberspace, "code is law," meaning that the way technology is designed (the architecture or code) can have as much of an impact on behaviour as traditional legal mechanisms. This is particularly salient in regulating AI, where the design of the technology itself can enforce privacy protections and other societal values (*ibid.*). For example, the use of algorithmic transparency and bias detection mechanisms in AI-driven financial services has gained attention in South Africa (Fundira, Edoun, Pradhan & Mbohwa, 2024). In a country with deep-seated socio-economic inequalities, the architecture of AI systems must ensure fairness and mitigate biases, especially in areas like credit scoring and employment recruitment, where discriminatory outcomes are likely (Fundira, Edoun, Pradhan & Mbohwa, 2024).

From a regulatory and policy standpoint in the African context, initiatives like Rwanda's National AI Policy are normative, as they seek to establish standards that align AI system development with national values and priorities (Rwanda Ministry of ICT and Innovation, 2020). This policy aims to shape the technical design of AI systems deployed in Rwanda, ensuring compatibility with local needs and ethical standards. A similar approach is evident in the national AI strategies of other African nations, including Nigeria, which also seek to

influence AI architecture to suit specific cultural and social contexts. These national efforts are complemented by continental initiatives such as the AUDA-NEPAD White Paper, *Regulation and Responsible Adoption of AI in Africa Towards Achievement of AU Agenda 2063* (AUDA-NEPAD, 2022), and the *Artificial Intelligence Roadmap for Africa: Contributing Towards A Continental AU Strategy On AI* (AUDA-NEPAD, 2022). These documents aim to develop shared guidelines for ethical AI development and use across Africa, fostering common norms and expectations around AI governance that respect cultural and social factors important to the continent's approach to technological regulation.

The call for ethical considerations in the development, use, and deployment of AI systems is quite resounding within various fields of AI[3]. As rational justifications for moral judgements, there is the need to underscore why they vary on how society needs to be organized. Ethics, as a construct, embodies several connotations and probabilities, most of which are 'good', but the concept is quite broad-ranging and challenging. This is because what is good will differ between a virtue ethicist and a deontologist, as well as between a utilitarian and a Confucian. It will also differ across borders. So, if policymakers seek a robust definition of ethical standards to serve as a benchmark for regulating AI and given the inherently cross-border nature of AI challenges, it becomes clear that an ideal definition would be universally accepted. Unfortunately, such a universally accepted definition currently does not exist.

But ethics are not enough. The value of creating ethical codes for AI has been widely undermined by the practices of 'ethics washing'[4] and 'ethics shopping'[5]. Critics have also provided overwhelming indications that the AI industry cannot be trusted with only voluntary ethical commitments (Karen Yeung, Andrew Howes, & Ganna Pogrebna, 2019). There needs to be mechanisms to reinforce normative claims, and these mechanisms have to go beyond self-governance and advance into concrete laws. Law and ethics are often complementary. However, not everything regulated is ethical, and not everything "ethical" is regulated. Law and ethics can be complementary – they can inform one another or indicate gaps.

In addition to the other modes of regulation, market forces are playing an increasingly significant regulatory role in jurisdictions like South Africa, where the government is exploring tax incentives for companies developing ethical AI systems (South African Department of Science and Innovation, 2021). This approach aims to create economic incentives for responsible AI development, aligning market forces with regulatory objectives.

---

[3] For example, when it comes to self-driving cars that use AI: Nyholm & Smids, 2016.

[4] Also called 'ethics theater', is the practice of fabricating or exaggerating an interest in equitable AI systems that work for everyone. 'A textbook example for tech giants is when a company promotes 'AI for good' initiatives with one hand while selling surveillance capitalism tech to governments and corporate customers with the other'. Khari Johnson, 'How AI companies can avoid ethics washing', July 2019

[5] 'Thus, is a world in which ethics-washing and ethics-shopping are becoming increasingly common, it is important to have common criteria based on which the quality of ethical and human rights commitments made can be evaluated. If not, there is a considerable danger such frameworks become arbitrary, optional, or meaningless rather than substantive, effective and rigorous ways to design technologies. When ethics are seen as an alternative to regulation, or as a substitute for fundamental rights, both ethics, rights and technology suffer.': Ben Wagner, 'Ethics as an escape from regulation: From ethics-washing to ethics-shopping' in Emre Bayamlioglu et al, eds.

Furthermore, Tunisia's AI strategy includes provisions for creating a market for AI safety research, illustrating how African nations are using economic tools to shape AI governance (Tunisia Ministry of Higher Education and Scientific Research, 2018). Similarly, Rwanda has positioned itself as an AI hub by offering incentives for AI companies and startups to establish operations within its borders, using economic incentives to attract AI innovation. This strategy allows Rwanda to influence the types of AI systems being developed and deployed, ensuring they align with national priorities. In Nigeria, the government has recently launched a ₦100 million AI Fund in partnership with Google to support startups in the AI sector, further demonstrating the use of market-based interventions to encourage local innovation and regulate AI development (TechPoint Africa, 2024). These efforts reflect how African countries are leveraging market forces to shape the AI landscape while aligning with broader regulatory goals. However, critics have raised concerns about whether market incentives alone can adequately govern AI, as they may prioritize profit over the ethical deployment of AI technologies. This reflects Lessig's caution about relying too heavily on markets as a sole regulatory modality, as it risks under-regulating areas where commercial interests conflict with public welfare.

17.3. Deciding the most appropriate regulation for governing AI.

If we examine past efforts to regulate emerging technologies in Africa, it becomes evident that policymakers have often demonstrated a fragile understanding of what regulation entails and how to effectively apply it to emerging technologies. This is evident in the tendency to apply existing regulatory frameworks to AI without considering its unique characteristics and potential impacts. However, experiences from other jurisdictions indicate that effectively regulating technologies like AI requires more than just the skillful application of traditional legal principles; it demands a new suite of regulatory responses that are not merely extensions of existing laws but are bespoke, tailored to the specific demands and complexities of local AI ecosystems.

There is the sentiment that AI is not just a technology that can be regulated by humans but one that can regulate humans, too – shaping humans in unanticipated ways – consciously and unconsciously, positively and negatively (Smuha, Nathalie A., 2021). Hence, because of this mutual influencing process, there has been some perplexity as to what form the regulation of AI should take, if at all it could be regulated.

The current perspectives on the regulation of AI cover various substantive and instrumental considerations. For example, the public interest approach to regulating AI situates the authorization of the technology in accordance with its public interest value (Alan Dignam, 2020). However, there is the critique that AI systems are not always developed with the public interest in mind (Karl Manheim & Lyric Kaplan, 2019).

Another perspective is that many technologies similar to AI have been (and are currently) regulated, so extant governance frameworks can extend to components of AI, and therefore, its regulation is a given (Miriam C Buiten, 2019). But there are still some ambiguities on whether to regulate the use of certain AI technologies, or instead, regulate where and what the technology is used to achieve (Sara Gerke et al, 2020). This last point is perhaps why policy proposals for regulating AI have resorted to doing so with principles of precaution, responsible innovation, permissionless innovation, and invention standards (Mimi S Afshar, 2022), amongst others (Oyeniyi Abe & Akinyi J Eurallyah, 2021).

Furthermore, AI regulation cannot be discussed outside the scope of the economic and market paradigms surrounding the technology[6]. Most AI systems are commercial and so there are liberalist notions that maintain that regulating AI can be overlaid on industry-specific actors and their parameters (Charles Kerrigan, 2022). But as a critique, if industry-specific actors and their parameters are relied on for regulating AI, they could be advancing the formations (and growth) of global capital at the expense of human rights by producing oversurveillance, targeting, stereotyping, bias and exclusion (Sonia K Katyal, 2019). Therefore extensive investigations of current legal reasoning could be less profitable approaches (Peter Wahlgren, 1984). Instead, proposals for regulating the technology could benefit from the formation of a more elaborate and critical philosophy that is truly interdisciplinary. The above considerations expose the complex intersections and perceived competition between law, science, technology, and the international community.

To answer the question of who should regulate the technology, approaches vary widely, but there are resounding propositions for an international and multi-stakeholder approach to regulating AI, which is in tandem with several established proposals made in the literature (H Güngör, 2020) and in the industry (María Belén Abdala, Andrés Ortega & Julia Pomares, 2020) (but not without its limitations) (Peter Cihon, Jonas Schuett & Seth D Baum, 2021). The multistakeholder proposal is mainly to establish trusted legal standards that will function as significant tools to steward the increasingly changing nature of AI systems and to ensure various approaches are considered to prevent harm from the technology, build trust, etc. (*ibid.*) In Africa, the issue of trust takes on particular significance. Widespread public mistrust in regulatory institutions means that efforts to regulate emerging technologies are often perceived as thinly disguised attempts to supress free speech[7]. The international-focused proposal is to achieve a shared responsibility from different countries and actors through collaboration, consultation, and coproduction[8] between states (perhaps one that involves new and dynamic systems-thinking, practice-oriented infrastructures, and even original empirical contributions) (Durose, Richardson, 2015, 1).

Pagallo and Bassi distinguish three kinds of legal regulation that could apply to AI systems (Ugo Pagallo & Eleonora Bassi, 2020). One is the traditional form of top-down regulation, such as legislation, that hinges on the threat of physical or pecuniary sanctions (*ibid.*, 441). Two is a bottom-up approach which is an assorted way of self-regulation that comes with

[6] Especially on issues surrounding trade, labor and antitrust (and consequences of AI in terms of employment, inequality, and competition). See Ajay Agrawal, Joshua Gans & Avi Goldfarb, "Economic policy for artificial intelligence" (2019) 19:1 Innovation policy and the economy 139–159.; Ajay Agrawal, Joshua Gans & Avi Goldfarb, *The economics of artificial intelligence: an agenda* (University of Chicago Press, 2019).

[7] "US TikTok ban "could embolden African governments" (March 15, 2024) African Business. <https://african.business/2024/03/technology-information/us-tiktok-ban-could-embolden-african-governments>

[8] Coproduction involves some form of "generative" discourses and social negotiations on how the development and deployment of AI should be guided by common grounds in terms of people, terminology, governance, and social values. Coproduction could open potential spaces for mediating the regulation of AI across states and could even involve a radically democratic alternative form of ethical design for the use of AI. For more on this type of generativity, see Runco and Albert, eds., *Theories of creativity*. Vol. 990. Newbury Park, CA: Sage, 1990.

legal constructions and limited accountability (*ibid.*). And then three is a form of co-regulation that could be exercised as what Pagallo et al. have elsewhere dubbed 'the middle-out interface' between top-down and bottom-up solutions between legislators and stakeholders (Ugo Pagallo, Pompeu Casanovas & Robert Madelin, 2019). This interface offers an alternative model of legal governance. It is now clear that the regulatory approaches towards the use of AI could be private, public, strict, flexible, domestic, international, ex-ante (forward-looking) or ex-post (backward-looking)[9].

17.4. **Addressing the challenges to regulating AI.**

Regulating a technology like AI presents a labyrinth of complexities, with jurisdictions worldwide grappling to formulate effective regulatory frameworks. Even the most advanced regions, such as the EU, have only recently managed to finalize regulations after years of deliberation and substantial resources[10]. For developing economies like those in Africa, the challenge of regulating AI is exacerbated by limited resources, public mistrust in regulatory institutions, and skepticism regarding policymakers' sincerity.

Defining what constitutes AI for regulatory purposes represents a crucial initial step, yet it remains a fundamental challenge. It is imperative to establish clear parameters delineating what falls within the scope of regulation and what does not. AI as an umbrella covers so many different things – from model-driven AI (rule-based) to data-driven AI (learning-based) and then robotics (hardware); the forms that the technology can take are unconstrained. This lack of definition is part of why people have fallen for 'Cheap AI' [11]– i.e., AI systems that are fundamentally rooted in pseudoscience[12], and also the trend of 'AI hype'[13] which is akin to

---

[9] The choice between ex-ante regulation (particularly regulatory standards) and ex-post regulation (particularly recalls and civil suits) implicates flexibility. Forward-looking rules may provide more certainty but less flexibility; backward-looking measures may provide more flexibility but less certainty. These trade-offs are particularly relevant to concerns raised about any liability from the use of AI systems. These concerns, however, likely derive at least as much from technical and algorithmic uncertainty (how will these AI systems actually perform) as from legal uncertainty (how will the law determine liability).

[10] European Parliament, "Artificial Intelligence Act: MEPs adopt landmark law" (March 13, 2024) Press Releases. <https://www.europarl.europa.eu/news/en/press-room/20240308IPR19015/artificial-intelligence-act-meps-adopt-landmark-law>

[11] Birhane defines Cheap AI, as 'a subset of Cheap science, [which] is produced when AI is inappropriately seen as a solution for challenges that it is not able to solve… Judgements made by these systems are inherently value-laden, wholly misguided and fundamentally rooted in pseudoscience.' See Abeba Birhane, 'Cheap AI' in Frederike Kaltheuner, (ed.), *Fake AI* (Meatspace Press, 2021) 41, at 42, 43, 46. See also the entire Fake AI book: https://fakeaibook.com, at 1 – 208). See also Shakir Mohamed, Marie-Therese Png, & William Isaac, 'Decolonial AI: Decolonial theory as sociotechnical foresight in artificial intelligence.' (2020) 33:4 *Philosophy & Technology* 659; Anna Bon et al. 'Decolonizing Technology and Society: A Perspective from the Global South' in Hannes Werthner, Erich Prem, Edward A. Lee, & Carlo Ghezzi (eds.), *Perspectives on Digital Humanism* (Springer, 2022) 61.

[12] 'Pseudoscience is where scientific claims are being made, but they're based on fundamentally shaky assumptions': Arvind Narayanan, 'AI snake oil, pseudoscience and hype: an interview with Arvind Narayanan' in Kaltheuner (ed.) (2021) 19, at 30.

[13] "Hype is exaggerated publicity which inflates expectations and prompts emotions such as excitement or fear, which can stop people in their tracks. It's a little like a magic show. By tugging on the emotions, hype causes

what Narayanan describes as 'AI snake oil'[14]– relating the dubious and exaggerated claims people make about AI capabilities (*ibid.*). More so, with the "AI effect"[15] what we consider AI changes over time because we get used to different things.

The impact of (real) AI can be broad, complex, and unpredictable. This is why regulators face challenges in fully understanding these impacts, particularly in areas like social behaviour and security. With the varying contexts where AI is applied (e.g. judicial systems, music recommendations, image generation, child welfare, etc.), regulators will need to determine whether to go for a horizontal approach (one AI regulation that applies all systems of AI) or a sectoral approach (one regulation per sector). While AI technology may be authorized or regulated for specific uses, it may also have many malicious uses that will become available. Therefore, separating intended and unintended uses will also be a challenge.

Africa's struggle with regulating AI mirrors its approach to other emerging technologies like social media. Some African countries have responded to calls for social media regulation by imposing outright bans due to concerns about inappropriate content undermining cultural and religious values, disregarding potential economic impacts[16]. This regulatory approach underscores a lack of sophistication that may extend to AI regulation.

AI is developing at a pace that often outstrips the ability of regulatory frameworks to keep up. This creates a lag between emerging AI applications and the establishment of appropriate regulations. Policymakers in Africa will have to decide if regulation should be technology-neutral or technology-specific (e.g. decide whether to prohibit surveillance in general or the use of AI for surveillance). There is also the consideration of creating unfair market advantages where if they prohibit a technology for one use, then they may be giving a competitive advantage to another (perhaps equally or more harmful technologies that are not AI). Most significant to this paper is the deterritorial nature of AI – where the technology transcends national borders, complicating regulation because of how the use of AI "dissolves borders" and demands action in borderless areas. Therefore, national laws may sometimes be

---

people to lend their ears, eyes, and brain waves to ideas, claims, and sales pitches in an uncritical manner.": Gemma Milne, 'Uses (and abuses) of Hype' in Kaltheuner (ed.) (2021) 115, at 115–124. See also Gemma Milne, Smoke & Mirrors: How Hype Obscures the Future and how to see past it (Robinson, 2020).

[14] "Much of what is sold commercially today as ''AI" is what I call "snake oil". We have no evidence that it works, and based on our scientific understanding of the relevant domains, we have strong reasons to believe that it couldn't possibly work… Some are not snake oil. There has been genuinely remarkable scientific progress. But because of this, companies put all kinds of systems under the AI umbrella—including those you would have more accurately called regression 20 years ago… But because of the hype, people have skipped this step and the public and policymakers have bought into it… There's this massive confusion around what AI is, which companies have exploited to create hype." Narayanan (2021) 20.

[15] The "AI effect" generally refers to a phenomenon in the field of artificial intelligence where once a problem is successfully solved by an AI system, the problem is no longer considered to be within the realm of AI. This effect highlights a moving goalpost for what is considered true artificial intelligence. The term encapsulates the idea that as AI technology advances and solves increasingly complex problems, those achievements are often reclassified as mere computing or automation rather than 'true' AI.

[16] "The rising cost of internet censorship in African countries" (January 13, 2024) TechCabal. <https://techcabal.com/2021/01/13/internet-censorship-in-africa/>

insufficient to regulate such a deterritorialized technology (Thomas Wischmeyer, & Timo Rademacher, eds., 2020).

Conclusion

With the proliferation of AI, the concept of regulation stands at a critical juncture, reflecting a complex interplay of how to govern such a technological innovation within existing legal considerations. This short paper attempted to explore the conceptual issues of regulation within AI governance by underscoring the unique characteristics of AI, including its deterritorialized nature and rapid evolution, and the demand for a supervisory framework that is both flexible and robust and capable of adapting to ongoing advancements while ensuring the maintenance of some core minimum standards. This process could entail mapping through the fragmented national, regional, and international policies to harmonize a cohesive framework that respects the diversity of legal systems yet strives for solidarity and a common ground in principles and practices. In essence, the future of AI regulation lies in an approach that balances innovation with responsibility but could leverage the lessons of international law-making from previous technological revolutions with new generative ideas toward effective governance.

References

Abdala, María Belén, Ortega, Andrés & Pomares, Julia. "Managing the transition to a multi-stakeholder artificial intelligence governance." G20 Insights, 2020. Available at: <https://www.g20-insights.org/policy_briefs/managing-the-transition-to-a-multi-stakeholder-artificial-intelligence-governance>.

Abe, Oyeniyi & Eurallyah, Akinyi J. "Regulating Artificial Intelligence through a Human Rights-Based Approach in Africa." African Journal of Legal Studies, 2021, vol. 14, no. 4, pp. 425–448.

Afshar, Mimi S. "Artificial Intelligence and Inventorship-Does the Patent Inventor Have to Be Human?" Hastings Science & Technology Law Journal, 2022, vol. 13, p. 55.

Agrawal, Ajay, Gans, Joshua & Goldfarb, Avi. "Economic policy for artificial intelligence." Innovation Policy and the Economy, 2019, vol. 19, no. 1, pp. 139–159.

Agrawal, Ajay, Gans, Joshua & Goldfarb, Avi. The Economics of Artificial Intelligence: An Agenda. University of Chicago Press, 2019.

Ben Wagner, 'Ethics as an escape from regulation: From ethics-washing to ethics-shopping' in Emre Bayamlioglu et al, eds.)

Birhane, Abeba. "Cheap AI." In Kaltheuner, Frederike (ed.), Fake AI. Meatspace Press, 2021, p. 41.

Black, Julia, and Dimity Kingsford Smith. "Critical reflections on regulation." Australasian Journal of Legal Philosophy, 2002, vol. 27, pp. 1-46.

Bon, Anna, et al. "Decolonizing Technology and Society: A Perspective from the Global South." In Werthner, Hannes, Prem, Erich, Lee, Edward A. & Ghezzi, Carlo (eds.), Perspectives on Digital Humanism. Springer, 2022, p. 61.

Buiten, Miriam C. "Towards intelligent regulation of artificial intelligence." European Journal of Risk Regulation, 2019, vol. 10, no. 1, pp. 41–59.

Cihon, Peter, Schuett, Jonas & Baum, Seth D. "Corporate governance of artificial intelligence in the public interest." Information, 2021, vol. 12, no. 7, p. 275.

Clarke, Roger. "Regulatory alternatives for AI." Computer Law & Security Review, 2019, vol. 35, no. 4, pp. 398-409.

Coeckelbergh, Mark. "Responsibility and the moral phenomenology of using self-driving cars." Applied Artificial Intelligence, 2016, vol. 30, no. 8, p. 748.

Didion, Joan. "Bureaucrats." In The White Album. New York: Simon and Schuster, 1979.

Dignam, Alan. "Artificial intelligence, tech corporate governance and the public interest regulatory response." Cambridge Journal of Regions, Economy and Society, 2020, vol. 13, no. 1, pp. 37–54.

Dohler, Marian. "Regulation." In Bevir, Mark (ed.), The Sage Handbook of Governance. Sage, 2011, pp. 518-534.

Durose, Catherine, and Richardson, Liz. Designing Public Policy for Co-Production: Theory, Practice and Change. Policy Press, 2015.

Egypt's Law Regulating and Developing the Use of Technology for Non-Banking Financial Activities (officially published in the Official Gazette on 8th of February 2022 and issued on the 14th of February 2022). Available at: <https://enterprise.press/stories/2022/01/09/egypt-now-has-a-regulatory-framework-for-the-fintech-industry-62179/>.

Egypt's Law Regulating and Developing the Use of Technology for Non-Banking Financial Activities, officially published in the Official Gazette on 8th of February 2022 and issued on the 14th of February 2022

European Parliament, "Artificial Intelligence Act: MEPs adopt landmark law" (March 13, 2024)

European Parliament. "Artificial Intelligence Act: MEPs adopt landmark law." Press Releases, 13 March 2024. Available at: <https://www.europarl.europa.eu/news/en/press-room/20240308IPR19015/artificial-intelligence-act-meps-adopt-landmark-law>.

Fundira, M., Edoun, E. I., Pradhan, A., & Mbohwa, C. (2024). Assessing digital competencies and AI ethics awareness among customers in the banking sector. *African Journal of Science, Technology, Innovation and Development*, 1–16.

Güngör, H. "Creating value with artificial intelligence: A multi-stakeholder perspective." Journal of Creating Value, 2020, vol. 6, no. 1, pp. 72–85.

Hill, Steven. "AI's Impact on Multilateral Military Cooperation: Experience from NATO." AJIL Unbound, 2020, vol. 114, pp. 147-151. Cambridge University Press.

Ifeanyi-Ajufo, Nnenna. "Digital Financial Inclusion and Security: The Regulation of Mobile Money in Ghana." Carnegie Endowment for International Peace, September 2022. Available at: <https://carnegieendowment.org/2022/09/19/digital-financial-inclusion-and-security-regulation-of-mobilemoney-in-ghana-pub-87949>.

Johnson, Khari. "How AI companies can avoid ethics washing." VentureBeat, July 2019. Available at: <https://venturebeat.com/2019/07/17/how-ai-companies-can-avoid-ethics-washing/>.

Kaltheuner, Frederike. Fake AI. Available at: <https://fakeaibook.com>.

Kerrigan, Charles. Artificial Intelligence: Law and Regulation. Edward Elgar Publishing, 2022, Cap. 6: Yee-Fen Lim, Hannah. "Regulatory Compliance," pp. 85-107.

Khari Johnson, 'How AI companies can avoid ethics washing', July 2019

Koop, Christel & Lodge, Martin. "What is regulation? An interdisciplinary concept analysis." Regulation & Governance, 2017, vol. 11, pp. 95-105.

Langford, Malcolm. "Taming the Digital Leviathan: Automated Decision-Making and International Human Rights." AJIL Unbound, 2020, vol. 114, pp. 141-146. Cambridge University Press.

Lessig, Lawrence. Code and Other Laws of Cyberspace. New York: Basic Books, 1999.

Lockey, Steven

Lockey, Steven et al. "A review of trust in artificial intelligence: Challenges, vulnerabilities and future directions." Information, 2021.

Manheim, Karl & Kaplan, Lyric. "Artificial intelligence: Risks to privacy and democracy." Yale Journal of Law & Technology, 2019, vol. 21, p. 106.

Millar, Jason. "An ethics evaluation tool for automating ethical decision-making in robots and self-driving cars." Applied Artificial Intelligence, 2016, vol. 30, no. 8, p. 787.

Milne, Gemma. Smoke & Mirrors: How Hype Obscures the Future and how to see past it. Robinson, 2020.

Narayanan, Arvind. "AI snake oil, pseudoscience and hype: an interview with Arvind Narayanan." In Kaltheuner, Frederike (ed.), Fake AI. 2021, p. 19, at p. 30.

Nyholm & Smids (2016); Borenstein, Herkert, & Miller (2019): 383-398; Sven Nyholm, 'The ethics of crashes with self-driving cars: a roadmap, II.' (2018) 13: e12506 Philosophy Compass 1; Jason Millar, 'An ethics evaluation tool for automating ethical decision-making in robots and self-driving cars' (2016) 30:8 Applied Artificial Intelligence 787; Mark Coeckelbergh, 'Responsibility and the moral phenomenology of using self-driving cars' (2016) 30:8 Applied Artificial Intelligence 748.

Orbach, Barak. "What Is Regulation?" Yale Journal on Regulation Online, 2012, vol. 30, p. 1.

Pagallo, Ugo & Bassi, Eleonora. "The Governance of Unmanned Aircraft Systems (UAS): Aviation Law, Human Rights, and the Free Movement of Data in the EU." Minds & Machines, 2020, vol. 30, p. 439.

Pagallo, Ugo, Casanovas, Pompeu & Madelin, Robert. "The middle-out approach: assessing models of legal governance in data protection, artificial intelligence, and the Web of Data." The Theory and Practice of Legislation, 2019, vol. 7, no. 1, p. 1.

Pereira, Luis Moniz, Santos, Francisco C. & Lenaerts, Tom. "To regulate or not: A social dynamics analysis of an idealised AI race." Journal of Artificial Intelligence Research, 2020, vol. 69, pp. 881–921.

Picker, Colin B. "A view from 40,000 feet: International law and the invisible hand of technology." Cardozo Law Review, 2001, vol. 23, p. 149.

Reed, Chris. "How should we regulate artificial intelligence?" Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 2018, vol. 376, no. 2128, p. 20170360.

Runco, Mark A., and Albert, Robert S. (eds.). *Theories of creativity*. Vol. 990. Newbury Park, CA: Sage, 1990.

Sara Gerke et al, "The need for a system view to regulate artificial intelligence/machine learning-based software as medical device", 2020

Scherer, Matthew U. "Regulating artificial intelligence systems: Risks, challenges, competencies, and strategies." Harvard Journal of Law & Technology, 2015, vol. 29, p. 353.

Singh, Divya. "Policing by Design: Artificial Intelligence, Predictive Policing and Human Rights in South Africa.", June 2022

Smith, Bryant Walker. "New Technologies and Old Treaties." AJIL Unbound 114, 2020

Smuha, Nathalie A., Beyond the Individual: Governing AI's Societal Harm, September 2021

Sonia K Katyal, "Private accountability in the age of artificial intelligence", 2019

Stix, Charlotte. "Actionable principles for artificial intelligence policy: three pathways." Science and Engineering Ethics, 2021, vol. 27, no. 1, pp. 1–17.

Tech Cabal, "The rising cost of internet censorship in African countries" (January 13, 2024) <https://techcabal.com/2021/01/13/internet-censorship-in-africa/>

The rising cost of internet censorship in African countries. TechCabal, 13 January 2024. Available at: <https://techcabal.com/2021/01/13/internet-censorship-in-africa/>.

US TikTok ban 'could embolden African governments. African Business, 15 March 2024. Available at: <https://african.business/2024/03/technology-information/us-tiktok-ban-could-embolden-african-governments>.

Wagner, Ben. "Ethics as an escape from regulation: From ethics-washing to ethics-shopping." In Bayamlioglu, Emre, et al (eds.), Being Profiled: Cogitas ergo sum: 10 years of profiling the European Citizen. 2018, p. 84, at p. 88.

Wahlgren, Peter. "A general theory of artificial intelligence and law." Legal Knowledge Based Systems JURIX, 1984, vol. 94, pp. 79–83.

Wilson, James Q. "The politics of regulation." In The Political Economy. Routledge, 2021, p. 82.

Wirtz, Bernd W., Weyerer, Jan C. & Sturm, Benjamin J. "The dark sides of artificial intelligence: An integrated AI governance framework for public administration." International Journal of Public Administration, 2020, vol. 43, no. 9, pp. 818–829.

Wischmeyer, Thomas & Rademacher, Timo (eds.). Regulating Artificial Intelligence. Springer, 2020.

Yeung, Karen, Howes, Andrew & Pogrebna, Ganna. "AI Governance by Human Rights-Centred Design, Deliberation and Oversight: An End to Ethics Washing." In Dubber, Markus D., Pasquale, Frank, & Das, Sunit (eds.), The Oxford Handbook of Ethics of AI. Oxford University Press, 2019, p. 76.

PART 5:
LATIN AMERICAN PERSPECTIVES ON AI GOVERNANCE

18. 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.

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 (The Future of AI in The Brazilian Judicial System., n.d.). 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 (DPRJ., 2022, July 21), 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 (DPRJ, 2020) 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 (*id.*). 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 (*id.*).

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 (*id.*).

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[1]. 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 can additionally provide insights towards better access to health in the state of Rio de Janeiro.

18.1. Justification

Brazil boasts the world's largest judicial system (*op. cit.*), both due to its core activity, which involves judicially safeguarding individual, collective, and social rights and due to its high rates of open data (Andrade, P., 2022). According to the Judges' Productivity Index (*idem*), 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.

[1] This data publicly available and updated on a yearly basis by the Brazilian National Council of Justice (CNJ), on the open platform *Justiça em Números* (Justice in Numbers), available at: https://justica-em-numeros.cnj.jus.br/ (Access on September 12, 2023).

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." (CNJ, 2022). 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 (Salomão, L. F., 2022). As an example, the development of the national platform for the management and training of AI models, Sinapses 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, 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.

18.2. 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 (Prevelakis, V., & Spinellis, D., 2001).

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 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.

18.3. 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.
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) (Google Patents). Also, there are Regulatory sandboxes in artificial intelligence such as the EU AI regulatory sandbox pilot program in Spain (Bläsing et al., 2010).	Operational Sandboxes are commonplace in the software development industry to debug and test software[2] and also in cybersecurity for malware detection[3].

[2] 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.

[3] 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.

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.

18.4. 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 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 (Leslie, D., Briggs, M., 2021). 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.

18.5. 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 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 start-ups, 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.

18.6. 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 (Noble, 2015). 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 (Leslie, D.,

2019). 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:

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; ● Diagnosis of a social (health) reality; ● Understanding the effectiveness of

	judicial intervention in terms of medication/treatment requisition; ● Scalability of the technology; ● Preserved procedural guarantees; ● Understanding the "map" of health needs not automatically met by current policies; ● Preserved procedural guarantees; ● Innovation; ● Understanding the personal situation of beneficiaries[4] 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; ● Does it affect the DPRJ's relationship with the judiciary? ● How to generate insights for the whole based on specific problems; ● Does it affect the speed of decision-making? ● Understanding of the tool by the employees; ● Visualisation of demand vs. necessity/priority; ● Does it require the digitization and standardisation of any work process? ● Difficulty in accessing data impacting development; ● What are we calling innovation? How do we measure the impact? ● Does AI technology promote

[4] 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.

	innovation in the judiciary system and the ecosystem? ● Impact on the rational use of medication;
3- What criteria can we use to analyse the impacts?	● Potential for discrimination, exclusion, or bias; ● Understanding of the tool by the employees; ● Inequity of judicialization; ● Speed of decision-making; ● Effective visualisation; ● Response time; ● Usability of the tool.
4- Can we mitigate potential negative impacts?	● Respecting the right to privacy governed by LGPD (Brazil's General Data Protection Law); ● Data anonymization; ● Security measures, such as encryption; ● Training administrative staff to use the tool; ● Indication that there might be demands not covered by the results provided by the developed technology; ● 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

18.7. Ethical Principles of AI Development

The misuse, abuse, or inadequate design of AI tools can lead to individual and societal harm (Leslie, D., 2019), 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, 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 concepts of FAIR principles[5] and the UNESCO guidelines for creating ethical AI (UNESCO, 2022) 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 ● Security and Protection ● Equity and Non-Discrimination ● Sustainability ● Right to Privacy and Data Protection ● Transparency and Explainability ● Responsibility and Accountability ● Awareness and Literacy ● Multistakeholder Governance and Collaboration ● Adaptive Promotion of Human Values
FAIR	● Findable ● Accessible ● Interoperable ● Interoperable

Under a dynamic session conducted with the Committee participants, it was asked of them to:

- Identify key existing ethical principles.

[5] 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 Wilkinson (2016); Go Fair (2017).

- 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;
- 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"[6].

18.8. 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 (Sambasivan, N., Arnesen, E., Hutchinson, B., Doshi, T., & Prabhakaran, V., 2021), 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

[6] 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.

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 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.

18.9. 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 (Cukierman, H. L., Teixeira, C., & Prikladnicki, R., 2007) . 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 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.

References

Andrade, P. (2022, September 5). Justiça em números 2022: cada magistrado julgou 6,3 processos por dia útil em 2021. AMB. Available at: https://www.amb.com.br/justica-em-numeros-2022-cada-magistrado-julgou-63-processos-por-dia-util-em-2021/

Bläsing, T.; Batyuk, L.; Schmidt, A. -D.; Camtepe, S. A.; and Albayrak, S. "An Android Application Sandbox system for suspicious software detection", 2010 5th International Conference on Malicious and Unwanted Software, Nancy, France, 2010, pp. 55-62, Available at: https://doi.org/10.1109/MALWARE.2010.5665792.

CNJ. 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. 2022. Available at: https://www.cnj.jus.br/wp-content/uploads/2022/09/seminario-e-justice-v6.pdf.

CNJ. Plataforma Sinapses. Available at: https://www.cnj.jus.br/sistemas/plataforma-sinapses/ (Access on September 12, 2023).

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

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. Available at: https://www.defensoria.rj.def.br/uploads/arquivos/09d3bcf2aa2c44e28fb55498d0a65f3d.pdf

DPRJ. (2022, July 21). Histórias do Plantão Noturno: defesa do direito à saúde é destaque. Available at: https://www.defensoria.rj.def.br/noticia/detalhes/20377-Historias-do-Plantao-Noturno-defesa-do-direito-a-saude-e-destaque

Four years and counting: what we've learned from regulatory sandboxes. Available at: https://blogs.worldbank.org/psd/four-years-and-counting-what-weve-learned-regulatory-sandboxes (Acesso em 12 de setembro de 2023)

GO FAIR. (2017). FAIR Principles - GO FAIR. Available at: https://www.go-fair.org/fair-principles/

IEEE Xplore. Available at: https://ieeexplore.ieee.org/abstract/document/5665792

Junquilho, Tainá Aguiar. Inteligência Artificial e Direito: limites éticos. Salvador: Juspodivm, 2022.

Justiça em Números (Justice in Numbers). Available at: https://justica-em-numeros.cnj.jus.br/ (Acesso em 12 de setembro de 2023).

Leslie, D. (2019). Understanding artificial intelligence ethics and safety: A guide for the responsible design and implementation of AI systems in the public sector. Public Policy Programme. Available at: https://doi.org/10.5281/zenodo.3240529

Leslie, D., & Briggs, M. (2021, March 20). Explaining decisions made with AI: A workbook (Use case 1: AI-assisted recruitment tool). ArXiv.org. Available at: https://arxiv.org/abs/2104.03906

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.

OECD. (2023). Regulatory sandboxes in artificial intelligence. Available at: https://www.oecd.org/sti/regulatory-sandboxes-in-artificial-intelligence-8f80a0e6-en.htm (Acesso em 12 de setembro de 2023)

Patent n. US8799862B2. Google Patents. https://patents.google.com/patent/US8799862B2/en

Plataforma Sinapses. Available at: https://www.cnj.jus.br/sistemas/plataforma-sinapses/ (Acesso em 12 de setembro de 2023)

Prevelakis, V., & Spinellis, D. (2001, June). Sandboxing Applications. In Usenix Annual Technical Conference, Freenix Track (pp. 119-126). Available at: https://www2.dmst.aueb.gr/dds/pubs/conf/2001-Freenix-Sandbox/html/sandbox32final.pdf

PubMed Central (PMC). Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4792175/ (Acesso em 12 de setembro de 2023)

Salomão, L. F. (2022). Artificial intelligence: technology applied to conflict management within the Brazilian Judiciary. Bibliotecadigital.fgv.br. Available at: https://bibliotecadigital.fgv.br/dspace/handle/10438/33954?locale-attribute=en

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

The Future of AI in The Brazilian Judicial System. ITS Rio. Available at: https://itsrio.org/en/publicacoes/the-future-of-ai-in-the-brazilian-judicial-system/ (Acesso em 24 de agosto de 2023).

UNESCO. (2022). Recommendation on the ethics of artificial intelligence. Available at: https://unesdoc.unesco.org/ark:/48223/pf0000381137

Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, Blomberg N, Boiten JW, da Silva Santos LB, Bourne PE, Bouwman J, Brookes AJ, Clark T, Crosas M, Dillo I, Dumon O, Edmunds S, Evelo CT, Finkers R, Gonzalez-Beltran A, Gray AJ, Groth P, Goble C, Grethe JS, Heringa J, 't Hoen PA, Hooft R, Kuhn T, Kok R, Kok J, Lusher SJ, Martone ME, Mons A, Packer AL, Persson B, Rocca-Serra P, Roos M, van Schaik R,

Sansone SA, Schultes E, Sengstag T, Slater T, Strawn G, Swertz MA, Thompson M, van der Lei J, van Mulligen E, Velterop J, Waagmeester A, Wittenburg P, Wolstencroft K, Zhao J, Mons B. *The FAIR Guiding Principles for scientific data management and stewardship*. Sci Data. 2016 Mar 15;3:160018. doi: 10.1038/sdata.2016.18. Erratum in: Sci Data. 2019 Mar 19;6(1):6. doi: 10.1038/s41597-019-0009-6.

19. 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.

Introduction

The growing use and development of artificial intelligence (AI)[1] has promoted a global race for regulatory frameworks, with the goal of developing so-called "trustworthy" AI systems (Smuha, N. A.., 2021). 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 (Collingridge, D., 1980). 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

[1] 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

influencing its social, political and innovation trajectories[2]. 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 (Gennus, A.; Stirling, A., 2017). Collingridge's approach emphasises active processes of learning from a particular class of past decisions in order to inform future decision-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 (Bromberg, L; Godwin A; Ramsay, I., 2017). To fill this ever-growing gap, a proposed solution is the development of new regulatory approaches that focus on public policy experimentation (Vermeulen, E.; Fenwick, M.; Kaal W. A., 2017).

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 (The Datasphere Initiative., 2022). 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 (World Bank., 2020). 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.

[2] 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/.

Over time, other regulators began to use the experimental regulation promoted by sandboxes in their own contexts. In Brazil[3], for example, in addition to the financial sector,[4] it is possible to identify initiatives in the telecommunications sector[5], in health[6], and in transport infrastructure[7]. 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) (According to Di Pietro, 2020), 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. 13.709/2018 (LGPD), the accountability principle[8] 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.[9] The proposals implemented by these

---

[3] 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.

[4] There are sandbox programs being conducted by three authorities of the National Financial System. For more details, see Part I of this article.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] This concept will be explained in Part II of this study.

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,[10] 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 behaviour of the regulateeB (Aranha, M. I., & Lopes, O. A., 2019) . However, given the paper's limitations , this analysis serves to start a debate which could be further developed in future studies.

19.1. 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 (Banco Central do Brasil - BACEN., 2020).

In the international context, the pioneer in the application of sandboxes was the British agency Financial Conduct Authority (FCA), in 2016 (World Bank., 2020). 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 (Organization for Economic Cooperation and Development - OECD., 2018). According to the institutions, their goals are to ensure innovation and business diversity, promoting competition and meeting user needs.

[10] 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.

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, including energy, transport and logistics infrastructure (BMWi - Federal Ministry for Economic Affairs and Energy., 2021) . 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 (European Securities and Markets Authority - ESMA; European Bank Authority - EBA; European Insurance And Occupational Pensions Authority - EIOPA., 2018) : (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 Start-ups Framework, Complementary Law No. 182/2021 (Brazil., 2021), 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 incidence of norms under their competence in relation to the regulated entity or to groups of regulated entities"[11]. For example, in ANATEL's proposed sandbox[12], 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.

[11] Author's translation.

[12] ANATEL's proposed sandbox was presented on items 369 to 372 of public consultation No. 41 of 2022 (ANATEL, 2022).

19.2. 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) (Personal Data Protection Commissioner – PDPC., 2017) ; the United Kingdom's Information Commissioner's Office – ICO (2019) (Information Commissioner's Office – ICO., 2019) ; Norway's Datatilsynet (2021) (Datatilsynet., 2021) ; Colombia's *Superintendencia de Industria y Comercio* – SIC, (2021) (Superintendencia De Industria Y Comercio – SIC, 2021) ; and France's *Commission Nationale de l'Informatique et des Libertés* – CNIL (2022) (Commission Nationale de l'Informatique et des Libertés – CNIL, 2022) . 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 (ANPD, 2023). The goals of the program are to allow participants to develop technologies that comply with personal data protection regulations, to be tested and analysed 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 (Gellert R., 2020) .

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 (Information Privacy Commissioner - IPC., 2018). 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 (Agencia Española De Protección De Datos – AEPD., 2019) (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 the privacy and protection of personal data (Moraes, T. et al., 2021). 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,[13] and the Brazilian LGPD.[14]

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.[15]

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.

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.

19.3. 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.

---

[13] 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".

[14] 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.

[15] 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.

The EU proposal, known as the AI Act (European Commission., 2021) , 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.[16] Furthermore, it empowers competent authorities[17] to establish a regulatory sandbox to "provide a controlled environment that facilitates the development, testing and validation of innovative AI 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 (Shimoda, C. A.; Moraes, T., 2023) . 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 (Brazil. Senado Federal Do Brasil., 2023) .

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[18], the Brazilian proposal establishes a rights protection regime. Thus, Bill n. 2,338/2023 proposes rights for individuals affected by AI 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

[16] 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"

[17] "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.

[18] 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.

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[19] 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

[19] Arts. 39 to 42 of the Bill.

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.

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) and Datatilsynet (2023) 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.

19.4. 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 (Grand View Research., 2022) . AI also fits into all the sectors mentioned in this study, and many more.

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.[20] 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.

[20] 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.

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 (Aranha, Lopes, 2019, op. cit.) . 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[21], or the potential use of regulatory sandboxes by SFN entities to experiment with blockchain and anticipate risks related to money laundering (Chagas, C., 2022). In addition, ANPD was inspired by this theory to develop its supervision and sanctioning administrative proceedings[22].

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) (Baptista, P.; Keller, C. I., 2016) 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, M., 2016) 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 (Ayres, I.; Braithwaite, J., 1992) responsive regulation, and Baldwin and Black's (2008) (Baldwin, R., & Black, J., 2008) truly responsive regulation.

Ayres and Braithwaite developed the theory of responsive regulation to transcend the impasse between "rigid" regulation and deregulation (Pereira, J. R., 2022) . In short, they intend to find the right balance between punishment and persuasion to make regulation effective. This

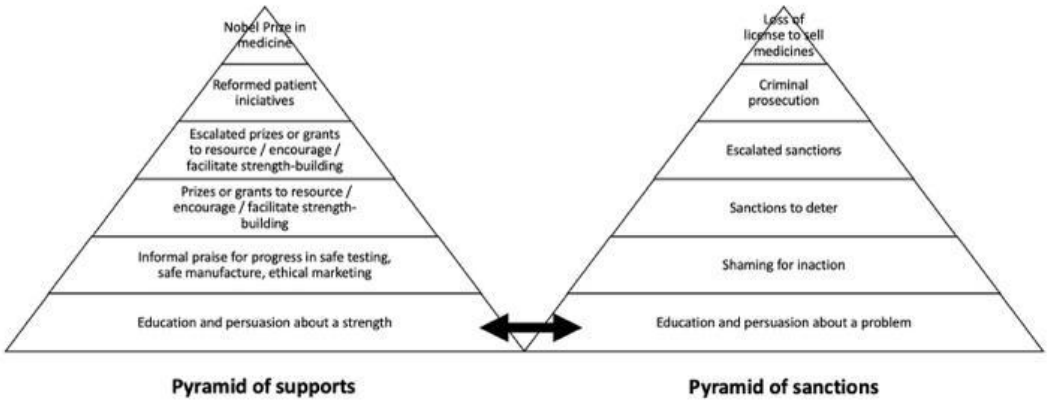
[21] In his paper, Goettenauer assesses whether there are elements of responsive regulation in the cybersecurity regulation of the SFN, presented in Resolution n. 4658, of April 26, 2018. The study concludes that it seems to be a new position by BACEN regarding the structuring of financial business in 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.

[22] 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.

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 behaviour (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) (*op. cit.*), based on Braithwaite, presented an example of such pyramids (see Figure 1).

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 awareness of the strengths and problems of a given business model. The persuasive approach can be seen 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) (*op. cit.*) suggest that the regulator, as a starting point, should always have cooperation in mind.

According to Baldwin, Cave and Lodge (2012) (Baldwin, R.; Cave, M.; Lodge, M., 2012) , 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) (*op. cit.*) 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 behaviour 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 behaviour 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 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 behaviour; (ii) *respond* to behaviour 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, J., 2021) 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 and inducing behaviour in regulated parties, either by encouraging compliance with the principles inherent in data protection legislation during testing, or by replicating behaviour 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.

19.5. 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 behaviour 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.

References

Agencia Española De Protección De Datos – AEPD. (2019). Guía De Privacidad Desde El Diseño, Madrid: AEPD. Available at: http://www.aepd.es/sites/default/files/2019-11/guia-privacidad-desde-diseno.pdf.

Agência Nacional de Saúde Suplementar – ANS. (2022). ANS promove webinar sobre SandBox Regulatório Prudencial, Brasília: ANS. Available at: https://www.gov.br/ans/pt-br/assuntos/noticias/periodo-eleitoral/ans-promove-webinar-sobre-sandbox-regulatorio-na-saude-suplementar.

Agência Nacional de Telecomunicações – ANATEL. (2022). Consulta pública nº 41, Brasília: ANATEL. Available at: https://apps.anatel.gov.br/ParticipaAnatel/VisualizarTextoConsulta.aspx?TelaDeOrigem=3&ConsultaId=10021.

Agência Nacional de Transportes Terrestres - ANTT. (2022). Resolução nº 5.999, de 3 de novembro de 2022. Brasília: ANTT. Available at: https://www.in.gov.br/en/web/dou/-/resolucao-n-5.999-de-3-de-novembro-de-2022-441284496.

ANATEL's proposed sandbox was presented on items 369 to 372 of public consultation No. 41 of 2022 (ANATEL, 2022).

ANPD. ANPD formaliza cooperação técnica com o Banco de Desenvolvimento da América Latina – CAF. 2023. Available at: https://www.gov.br/anpd/pt-br/assuntos/noticias/anpd-formaliza-cooperacao-tecnica-com-o-banco-de-desenvolvimento-da-america-latina-2013-caf

Aranha, M. I., & Lopes, O. A. (2019). 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). Brasília: Centro de Políticas, Direito, Economia e Tecnologias das Comunicações da UnB.

Ayres, I., & Braithwaite, J. (1992). Responsive regulation: Transcending the deregulation debate. Oxford Walton Street: Oxford University Press.

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

Baldwin, R., Cave, M., & Lodge, M. (2012). Understanding regulation: theory, strategy, and practice. Oxford: Oxford University Press, 2 ed.

Banco Central do Brasil - BACEN. (2020). Fintechs. Brasília: BACEN. Available at: https://www.bcb.gov.br/estabilidadefinanceira/fintechs.

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.

Black, J. (2021). Constitutionalising Regulatory Governance Systems, LSE Law, Society and Economy Working Papers. Available at: https://eprints.lse.ac.uk/113670/1/Black_constitutionalising_regulatory_governance_published.pdf

BMWi - Federal Ministry for Economic Affairs and Energy. (2021). Making space for innovation: The handbook for regulatory sandboxes. Available at: https://www.bmwk.de/Redaktion/EN/Publikationen/Digitale-Welt/handbook-regulatory-sandboxes.pdf%3F__blob%3DpublicationFile%26v%3D2

Brazil. (2021). Lei Complementar nº 182/2021, de 1º de junho de 2021. Marco Legal das Startups. Available at: https://www.planalto.gov.br/ccivil_03/LEIS/LCP/Lcp182.htm

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. Available at: https://www25.senado.leg.br/web/atividade/materias/-/materia/157233

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). Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3090844

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. Available at: https://pos.direito.ufmg.br/downloads/Novas-fronteiras-do-sistema-financeiro-nacional.pdf

Collingridge, D. (1980). The Social Control of Technology. Frances Pinter Publisher Ltd, 16.

Commission Nationale de l'Informatique et des Libertés – CNIL. (2022). EdTech "sandbox": the CNIL supports ten innovative projects. Paris: CNIL. Available at: https://www.cnil.fr/en/edtech-sandbox-cnil-supports-10-innovative-projects

Datatilsynet. (2021). Sandbox for responsible artificial intelligence, Oslo: Datatilsynet. Available at: https://www.datatilsynet.no/en/regulations-and-tools/sandbox-for-artificial-intelligence/

Datatilsynet. (2023). Reports. Oslo: Datatilsynet. Available at: https://www.datatilsynet.no/en/regulations-and-tools/sandbox-for-artificial-intelligence/reports/

Di Pietro, M. S. Z. Direito administrativo. 3 ed. Rio de Janeiro: Forense, 2020.

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. Available at: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52021PC0206

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. Available at: https://www.europarl.europa.eu/doceo/document/TA-9-2023-0236_EN.pdf

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. Available at: https://www.esma.europa.eu/sites/default/files/library/jc_2018_74_joint_report_on_regulatory_sandboxes_and_innovation_hubs.pdf

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". Available at: 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

Gellert, R. (2020). The Risk-Based Approach To Data Protection, Oxford Scholarship Online, Oxford: Oxford University Press. Available at: https://academic.oup.com/book/40487

Gennus, A., & Stirling, A. (2017). Collingridge and the dilemma of control: Towards responsible and accountable innovation. Research Policy. Available at: http://dx.doi.org/10.1016/j.respol.2017.09.012

Goettenauer, C. (2019). Regulação Responsiva e a Política de Segurança Cibernética do Sistema Financeiro Nacional. Journal of Law and Regulation, 5(1), 131–146. Available at: https://periodicos.unb.br/index.php/rdsr/article/view/20944

Grand View Research. (2022). Artificial Intelligence In Fintech Market Size, Share &

Hersey, F. (2023). EU AI Act definition of AI aligns with OECD definition, biometric risk updated. Biometric Update.com. Available at: https://www.biometricupdate.com/202303/eu-ai-act-definition-of-ai-aligns-with-oecd-definition-biometric-risk-updated

ICO. Regulatory Sandbox – Previous Participants. London: ICO. 2023. Available at: https://ico.org.uk/for-organisations/regulatory-sandbox/previous-participants/

Information Commissioner's Office – ICO. (2019). Regulatory Sandbox. London: ICO. Available at: https://ico.org.uk/sandbox

Information Privacy Commissioner - IPC. (2018). Privacy by Design. Toronto: IPC. Available at: https://www.ipc.on.ca/wp-content/uploads/2018/01/pbd.pdf

Justiça em Números (Justice in Numbers), available at: https://justica-em-numeros.cnj.jus.br/ (Access on September 12, 2023).

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. Available at: http://dx.doi.org/10.1093/idpl/ipaa025

Organization for Economic Cooperation and Development - OECD. (2018). Financial Markets, Insurance and Private Pensions: Digitalisation and Finance. Paris: OECD. Available at: https://www.oecd.org/finance/private-pensions/Financial-markets-insurance-pensions-digitalisation-and-finance.pdf

Pereira, J. R. (2022). Openness Doesn't Hurt: Enforcing Qualified Machine-Learning Transparency For Data Protection Through Responsive Regulation. Brasília: Universidade de Brasilia.

Personal Data Protection Commissioner – PDPC. (2017). A Trusted Ecosystem for Data Innovation. Singapore: PDPC. Available at: https://www.pdpc.gov.sg/-/media/Files/PDPC/New_DPO_Connect/aug_2017/pdf/ATrustedEcosystemForDataInnovation.pdf

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. Available at: https://www.in.gov.br/en/web/dou/-/resolucao-cd/anpd-n-1-de-28-de-outubro-de-2021-355817513

Rio de Janeiro. (2022). Sandbox.rio. Available at: https://www.sandboxrio.com.br/sobre.html

Shimoda, C. A., & Moraes, T. (2023). Brazil's Path to responsible AI. OECD.AI Wonk Blog. Available at: https://oecd.ai/en/wonk/brazils-path-to-responsible-ai

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. Available at: https://ssrn.com/abstract=3501410

Superintendencia De Industria Y Comercio – SIC. (2021). Sandbox on privacy by design and by default in Artificial Intelligence, Bogotá: SIC. Available at: https://www.sic.gov.co/sites/default/files/files/2021/150421%20Sandbox%20on%20privacy%20by%20design%20and%20by%20default%20in%20AI%20projects.pdf

The Datasphere Initiative. (2022). Sandboxes for data: creating spaces for agile solutions across borders. Available at: https://www.thedatasphere.org/datasphere-publish/sandboxes-for-data/

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. Available at: https://www.grandviewresearch.com/industry-analysis/artificial-intelligence-in-fintech-market-report

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). Available at: http://dx.doi.org/10.2139/ssrn.2834531

Wansley, M. (2016). Regulation of emerging risks. Vanderbilt Law Review, 69(401). Available at: http://ssrn.com/abstract=2646316

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. Available

at: https://www.estadao.com.br/politica/pressao-e-ameaca-no-congresso-como-o-google-derrubou-o-pl-2630-das-fake-news-em-14-dias/

World Bank. (2020). Global Experiences from Regulatory Sandboxes. Washington, DC: World Bank. Available at: https://openknowledge.worldbank.org/handle/10986/34789

Brasil, Lei n°13.709/2018 (LGPD).

Resolution CD/ANPD n. 1, of October 28, 2021.

Information Privacy Commissioner - IPC., 2018

Pre-print version of Belli L. Gaspar. W.B. The Quest for AI Sovereignty, Transparency and Accountability. Springer-Nature. (2025).

20. Building a repository of public algorithms: Case study of the dataset on automated decision-making systems in the Colombian public sector

Juan David Gutiérrez[1] and Sarah Muñoz-Cadena[2]

Abstract

This chapter 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 chapter also explains the main difficulties that the researchers encountered as well as the solutions that were implemented. Finally, the chapter 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.

Introduction[3]

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 (Juan David Gutiérrez, Sarah Muñoz-Cadena, 2023). The apparent under-registration of 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, 2023).

[1] Associate professor at Universidad de los Andes. Corresponding author. Email: juagutie@uniandes.edu.co.

[2] Researcher at Policéntrico and master's student at Universidad del Rosario. Email: sarahm.munoz@urosario.edu.co.

[3] 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 the two authors constructed along with the research assistance of Michelle Castellanos-Sánchez and some of the descriptive statistics were included in a chapter recently published by the authors.

This chapter 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 chapter also explains the main difficulties that the researchers encountered as well as the solutions that were implemented. Finally, the chapter reflects on the challenges of fostering algorithmic transparency in a Global South country and offers recommendations for replicating this project in other countries.

This chapter 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.

20.1. 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')" (Meeri Haataja, Linda van de Fliert, Pasi Rautio, 2020) and to understand where the data comes from and what results the system may produce ('traceability').

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.[4] The following table summarizes sixteen repositories that are publicly available online, including the new dataset documented in this chapter:

Table 1 – Repositories of public algorithms[5]

Name of the repository	Geographical Scope	Organization that published it	Nature of the organization	# of registered systems
America				
Observatorio Algorítmico	Argentina	Algorithmic Avengers	Civil society organization	10
Algoritmos públicos	Chile	GobLab – Universidad Adolfo Ibáñez	University	75

[4] 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/

[5] With information updated on November 27, 2023.

Tablero de seguimiento marco ético[6]	Colombia	National government	Public	6
Portal de Datos Abiertos	Colombia	National government	Public	16
Ejercicios de Innovación Basados en Inteligencia Artificial	Colombia	National government	Public	6
Sistemas de decisión automatizada en el sector público colombiano	Colombia	Universidad del Rosario	University	113
Algorithmic tools	New York, United States	Subnational government	Public	31
Artificial Intelligence and Algorithms	Ontario, Canada	Regional government	Public	8
Algorithm tips	United States	Northwestern University	University	903
Proyectos de aplicación de Inteligencia Artificial	Uruguay	National government	Public	2
Europe				
City of Amsterdam Algorithm Register	Amsterdam, The Netherlands	Subnational government	Public	4
Inventaire des algorithmes utilisés par la Ville d'Antibes	Antibes Juan-les-Pins, France	Subnational government	Public	8
Publication des algorithmes et des codes sources	France	National government	Public	14
Artificial intelligence systems of Helsinki	Helsinki, Finland	Subnational government	Public	9
Consultation des Algorithmes publics de Nantes Métropole	Nantes, France	Subnational government	Public	2
EU Selected AI cases in the public sector - AI-WATCH: EU	European Union	Supranational government	Public	686

[6] Since May 2023 the information is no longer available.

Artificial Intelligence Observatory				
AI-X: AI Public Services Explorer	European Union	Supranational government	Public	143
Algorithmic Transparency Reports	United Kingdom	National government	Public	6
Global				
Observatory of Algorithms with Social Impact - OASI	Worldwide	Eticas Foundation	Civil society organization	152

Source: Authors' own elaboration based on Gutiérrez (Foro GPP, 2023)

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 (Romina., José Pablo L., María Paz H., 2021).

The research that produced the new dataset for the Colombian public sector, documented in this chapter, 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.

20.2. How we built our dataset of ADMs in the Colombian public sector

20.2.1. 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*).[7] 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.[8]

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[9], hence number of entities that used ADMs would probably be lower.

Our new database confirmed the under-registration of algorithms in the government-based repositories, as we will explain in the following pages.

20.2.2. 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

[7] 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

[8] The FURAG response data can be consulted on the following platform: https://www.datos.gov.co/Funci-n-p-blica/FURAG/daed-z4fw/data

[9] The June 2020 glossary (version 5) of the FURAG is available at the following URL: https://www.funcionpublica.gov.co/documents/28587410/36200637/Glosario_mipg.pdf/9ff42c08-61a9-e0fa-76b1-1f662c0b2202?t=1593207412671

software and antivirus software)[10]. 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 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.

[10] Defining what an algorithm is and which one has certain characteristics and should therefore be notified is not a minor issue. In fact, in a reflection on the French experience in the construction of public registers of algorithms, the first conclusion is that "defining algorithms can be challenging" (Pénicaud, 2021).

The database characterizes each system with regards to 40 variables which can grouped into five categories: (i) basic information on the ADM systems, 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 ADM systems 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 (Juan David Gutiérrez, 2023).

20.3. Main findings

Although this chapter 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.

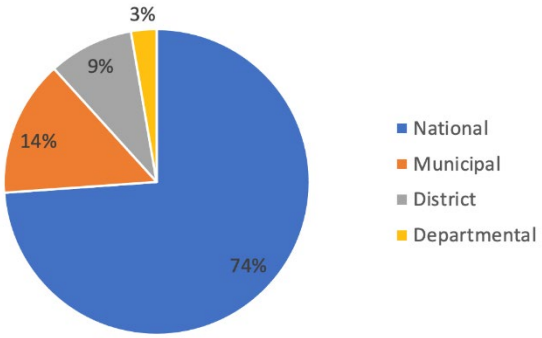
The public entities mapped in this research that are using ADMs in Colombia are, for the most part, from the executive branch (93%) and, to a lesser extent, from the judicial branch (4%) or belong to control entities (3%). We did not find any system adopted by the legislative branch.

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).[11] In the case of local entities, these are

[11] 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.

distributed among initiatives led by municipal (14%), district (9%) and departmental (3%) entities (see Figure 1).

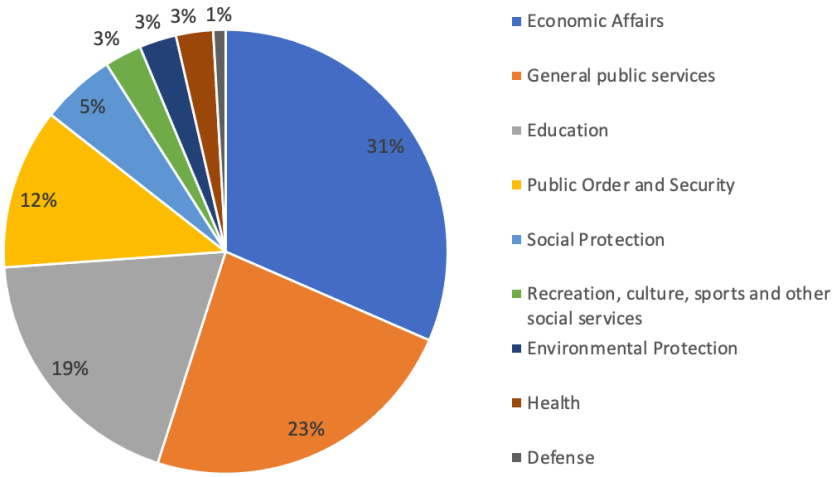
Figure 1. Percentage of systems adopted by national or territorial entities.



Source: Authors' own elaboration

Applying the Classification of the Functions of Government (COFOG)[12], we found that 74% of the mapped systems were adopted by public entities that perform three categories of functions: "economic affairs" (32%), "general public services" (23%) and "education" (19%) (see Figure 2).

Figure 2. ADMs classification based on the first level of COFOG.



Source: Authors' own elaboration

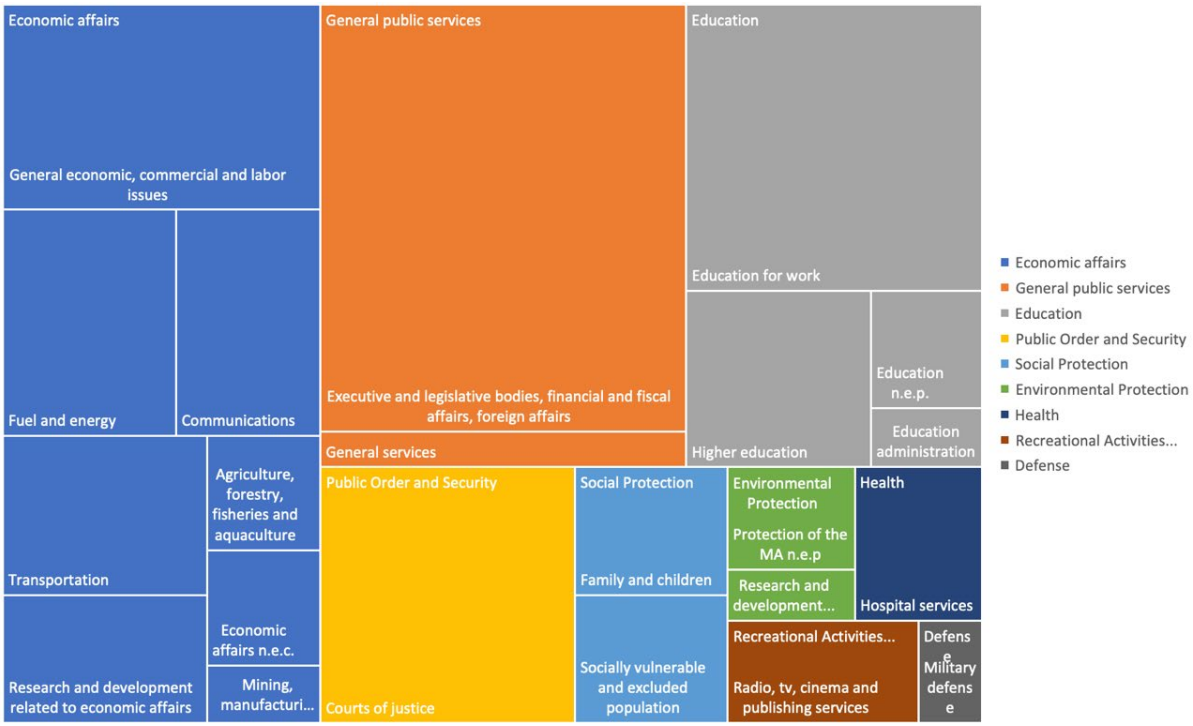
This finding for the Colombian case contrasts with that reported in the GobLab UAI Public Algorithms Repository: in the Chilean case, of the 75 algorithms that are registered (GobLab UAI, 2023, 50), 61% are related to three sectors: "health" (25%), "economic affairs" (24%) and "public order and security" (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

[12] "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 accordance with the purposes or public functions, showing the nature of the services provided by the institutions on behalf of the state" (Dane, 2020).

the three categories with the most mapped algorithms were: “general public services”, “economic affairs”, and “public order and safety” (Gianluca M., Colin Van N. Anis B., 2020).

Returning to our database, COFOG allows us to detail the type of function performed by the public entities that adopted ADMs 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 ADMs.



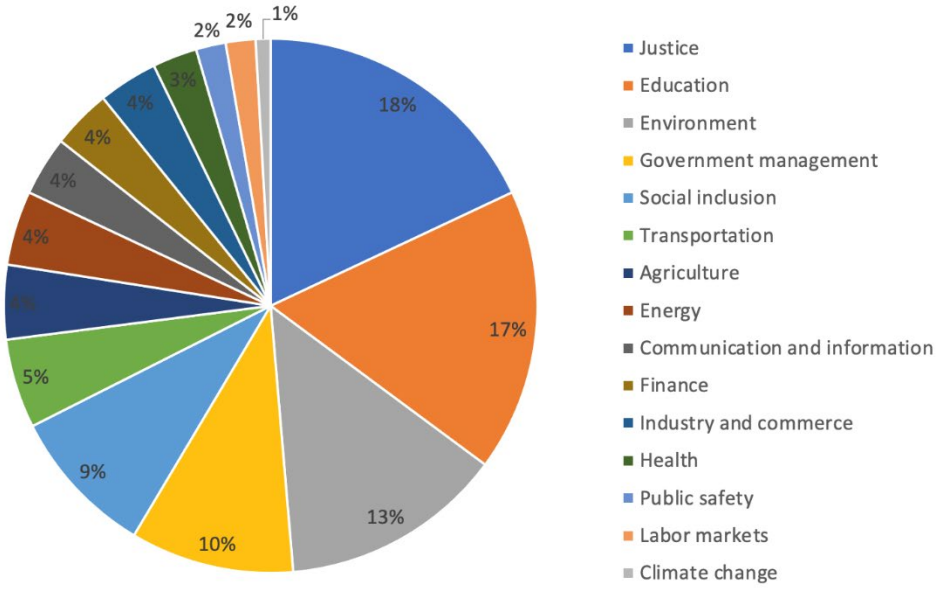
Source: Authors’ own elaboration

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 (Figure 4).[13] 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

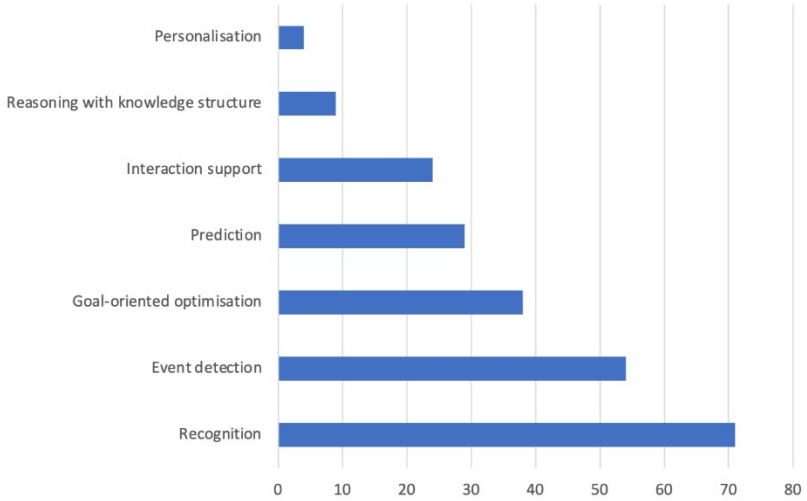


Source: Authors' own elaboration

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 (OECD, 2022). 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 (8%) perform knowledge-structured reasoning and four (4%) seek to personalize (see Figure 5).

[13] 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.

Figure 5. Classification of ADMs in Colombia by type of functionalities



Source: Authors' own elaboration

20.4. Reflections on building repositories of public algorithms

This chapter 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 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 based on more than 300 sources of primary and secondary information.

We would like to close this chapter 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[14]. What in the emerging literature on the subject is known as "registry algorithmic transparency" (José Pablo L., Romina G. and María Paz H., 2023). 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

[14] For a discussion on the links between Open Government policies and algorithmic transparency in Colombia, see Gutiérrez and Castellanos-Sánchez (2023).

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. Because as mentioned by Soizic Pénicaud: "Public algorithms are at a crossroads between open data, data protection, open source and access to administrative documents" (Pénicaud, 2021) and hence collaboration between different public and private entities might help in this complex balance.

On one hand, 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 it is being used, but also to know how the system was built (e. g. the data used to train, if it is 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[15]. Not being able to access such information makes impossible for third parties to assess the performance of the system.

In addition, the information made available to the public about the systems must be in clear and simple language or for "specific audiences" (Ada Lovelace Institute et al., 2021), because although "most of the information lies in the hands of domain-experts using the tools" (Pénicaud, 2021), the automation of certain decisions based on the use of these tools can directly affect the lives of citizens (and even their rights), in some cases those of the most vulnerable (Lapostol et al., 2023). Such is the case of algorithms that make it possible to determine whether a citizen is a beneficiary of a government policy or program.

On the other hand, 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.

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

[15] 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.

new systems may be implemented.[16] 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 updated database in 2024.

References

ADA LOVELACE INSTITUTE; AI NOW INSTITUTE; OPEN GOVERNMENT PARTNERSHIP. Algorithmic accountability for the public sector. Learning from the first wave of policy implementation. 2021.

Colombia. Medición del Desempeño Institucional. Función Pública – Gov.CO. Available at: https://www1.funcionpublica.gov.co/web/mipg/medicion_desempeno.

DANE. Clasificación de las funciones del gobierno (COFOG). 2020. Available at: <https://www.dane.gov.co/files/sen/nomenclatura/cofog/COFOG-AC.pdf>.

Foro Administración, Gestión y Política Pública (Foro GPP). Repositorios y registros públicos de algoritmos. 2023. Available at: https://forogpp.com/inteligencia-artificial/repositorios-y-registros-de-algoritmos/.

GARRIDO, R.; LAPOSTOL, J. P.; HERMOSILLA, M. P. Transparencia algorítmica en el sector público. Santiago de Chile: GOB LAB UAI. Consejo para la Transparencia, out. 2021. Available at: <https://goblab.uai.cl/wp-content/uploads/2021/10/ESTUDIO-TRANSPARENCIA-ALGORITMICA-EN-EL-SECTOR-PUBLICO-GOBLAB-CPLT-final....pdf>.

GOBLAB UAI. Repositorio Algoritmos Públicos. Informe Anual 2023. Santiago de Chile: Universidad Adolfo Ibáñez (UAI), mar. 2023.

GOBLAB UAI. Repositorio de algoritmos públicos de Chile. Primer informe de estado de uso de algoritmos en el sector público. [s.l.] Universidad Adolfo Ibáñez (UAI), 2022. Available at: <https://goblab.uai.cl/wp-content/uploads/2022/02/Primer-Informe-Repositorio-Algoritmos-Publicos-en-Chile.pdf>.

GUTIÉRREZ, J. D. ¿Están los jueces en capacidad de usar modelos de lenguaje a gran escala (LLMs)? Revista EXCEJLENCIA, v. 7, p. 10–15, maio 2023c.

GUTIÉRREZ, J. D. ChatGPT in Colombian Courts: Why we need to have a conversation about the digital literacy of the judiciary. VerfBlog, 23 fev. 2023a. Available at: <https://verfassungsblog.de/colombian-chatgpt/>

GUTIÉRREZ, J. D. Hablemos sobre el uso de ChatGPT para redactar decisiones judiciales. La Silla Vacía, 26 fev. 2023b. Available at: <https://juangutierrez.co/publicaciones/otras-

[16] Keeping the information updated is a great challenge because, for example, one of the systems that was active when the database was built, in November 2023 was suspended by the entity that implemented it after an internal control audit of the system.

publicaciones/#:~:text=Hablemos%20sobre%20el%20uso%20de%20ChatGPT%20para%20redactar%20decisiones%20judiciales>

GUTIÉRREZ, J. D. Repositorios y registros públicos de algoritmos. Foro Administración, Gestión y Política Pública, 6 jul. 2023d. Available at: <https://forogpp.com/inteligencia-artificial/repositorios-y-registros-de-algoritmos/>. Acesso em: 26 jul. 2023.

GUTIÉRREZ, J. D.; CASTELLANOS-SÁNCHEZ, M. Transparencia algorítmica y Estado Abierto en Colombia. Revista Reflexión Política, v. 25, n. 52, 2023.

GUTIÉRREZ, J. D.; MUÑOZ-CADENA, S. Adopción de sistemas de decisión automatizada en el sector público: Cartografía de 113 sistemas en Colombia. GIGAPP Estudios Working Papers, v. 10, n. 270, p. 365–395, 25 set. 2023.

GUTIÉRREZ, J. D.; MUÑOZ-CADENA, S.; CASTELLANOS-SÁNCHEZ, M. Sistemas de decisión automatizada en el sector público colombiano [Dataset]. Universidad del Rosario, 2023. Available at: <https://doi.org/10.34848/YN1CRT>

HAATAJA, M.; VAN DE FLIERT, L.; RAUTIO, P. Public AI Registers. Realising AI transparency and civic participation in government use of AI. [s.l: s.n.]. Available at: <https://algoritmeregister.amsterdam.nl/wp-content/uploads/White-Paper.pdf>.

LAPOSTOL, J. P.; GARRIDO, R.; HERMOSILLA, M. P. Algorithmic Transparency from the South: Examining the state of algorithmic transparency in Chile's public administration algorithms. Em: 2023 ACM CONFERENCE ON FAIRNESS, ACCOUNTABILITY, AND TRANS-PARENCY (FACCT '23). Chicago, IL, USA: ACM, 2023. Available at: <https://doi.org/10.1145/3593013.3593991>

MISURACA, G.; VAN NOORDT, C.; BOUKLI, A. The use of AI in public services: results from a preliminary mapping across the EU. Em: PROCEEDINGS OF THE 13TH INTERNATIONAL CONFERENCE ON THEORY AND PRACTICE OF ELECTRONIC GOVERNANCE (ICEGOV 2020). Athens, Greece: 23 set. 2020.

OECD. OECD Framework for the Classification of AI systems. Paris: OECD, 22 fev. 2022. Available at: <https://www.oecd-ilibrary.org/science-and-technology/oecd-framework-for-the-classification-of-ai-systems_cb6d9eca-en>. Acesso em: 2 jun. 2022.

PÉNICAUD, S. Building Public Algorithm Registers: Lessons Learned from the French Approach. Open Government Partnership, 12 maio 2021. Available at: https://www.opengovpartnership.org/stories/building-public-algorithm-registers-lessons-learned-from-the-french-approach/

21. International efforts aimed at promoting AI transparency and/or accountability

"If a machine is expected to be infallible, it cannot be intelligent either."

- Alan Turing

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.

Abstract

Artificial Intelligence (AI) assumes benefits but also sets out challenges for its regulation, so it is essential to know up to date which international efforts have been made to regulate it, highlighting the importance of promoting the principles of transparency and accountability in the creation and implementation of technologies that use artificial intelligence systems. In addition, the landscape of responsible governance of AI will be displayed to generate trust in users, through the implementation of its principles both nationally and internationally, to maximize the benefits for society and minimize potential risks of its use, through collaboration, among countries and multi-stakeholders for the promotion of trustworthy and ethical AI.

The final approach of this document focuses on the presentation of a Latin American proposal based on existing international jurisprudence aimed at the creation of an ex profeso mechanism that contributes to matters related to artificial intelligence in this region, through cooperation and the establishment of strategic alliances with international organizations such as the Organization of American States (OAS), through the Inter-American Juridical Committee, and with the support and participation of other economic blocks at a global scale that have shown their interest in the topic.

Introduction

International efforts aimed at regulating artificial intelligence (AI) are increasingly important in an interconnected and technology-dependent world. Various international organizations and coalitions have emerged and come together to establish guidelines and regulations that guide the development and application of AI responsibly to improve their standards and regulatory frameworks that promote transparency, impartiality, privacy, and inclusion in their processes.

As technological evolution accelerates, we have seen societies and economies become stronger by improving living conditions and increasing productivity, so we must highlight the importance of addressing common global technological governance challenges to identify possible gaps without stopping the course of innovation.

For the above, this work has an approach based on the theory of multilateralism, bearing in mind the international dynamics that have contributed to the transformation experienced in recent years, as a result of the advances brought by technological innovation that has impacted the local, regional and global spheres that positively or negatively affect our societies.

By its definition, we understand that multilateralism is conceived as "a series of transitional arrangements between more than two States that, having found points of common interest, propose to transform them into collective objectives and actions" (Cox, 1996).

In addition to the conception of States to which this theory refers, over time, actors on the international scene were added who have played a key role in the development and implementation of international cooperation and effective multilateral practices, such as international organizations and organized civil society.

We can see the above reflected in the impetus and proactivity shown by various international organizations that have promoted specific guidelines and orientations for the responsible use of artificial intelligence systems, always with strict respect and adherence to human rights with a human-centric approach. An example of the foregoing are the exercises that have been launched such as the Council Recommendation on Artificial Intelligence issued by the Organization for Economic Cooperation and Development (OECD); the Recommendation on the Ethics of Artificial Intelligence of the United Nations Educational, Scientific and Cultural Organization (UNESCO); the Artificial Intelligence Law of the European Union and the groups of world economic powers that have pronounced in this regard through the G7 Hiroshima Leaders' Communiqué and the Ministerial Declaration of the G7 Technology; and Digital Ministers' Meeting, as well as the communiqué of the G20 commitment group.

21.1. International efforts aimed at regulating AI

International collaboration is essential for AI since it knows no borders and its impact extends to multiple sectors, from the health sector to issues related to the environment. With the purpose of ensuring equitable and ethical development of AI, countries, and regions must exchange knowledge and good practices. This must be done in accordance with international law and promoting cooperation among countries.

Regulatory efforts up to date are insufficient and there is no current regulation on the matter in charge of monitoring and preventing latent risks, despite the existing proposal for an AI law in Europe, which is estimated to come into force in 2026 (Ziady Hanna, 2023) (CNN Business, 2023). None of the recent initiatives that come from international organizations are binding, so their scope and impact are flexible and voluntary under the figure of recommendations or principles that do not guarantee the success of AI governance.

Even though, these exercises are part of the first set of international regulatory instruments to face the challenges posed by AI, through a suggested governance framework that provides solutions focused on global stability and technological balance.

20.1.1. OCDE

The first global instrument was “The Council Recommendation on Artificial Intelligence” adopted in May 2019 (OECD, 2019). The recommendation establishes practical and flexible principles to remain in force overtime and that complement other standards such as privacy, digital security risk management, and responsible business conduct.

20.1.2.G20

Subsequently, in June 2019, the G20 adopted the Human-Centered AI Principles that are based on the OECD AI Principles, broadly stating that considering how technology affects society, a favorable environment must be provided for human-centered development.

This paper shows that AI technologies can help promote inclusive economic growth and bring great benefits to society; Its responsible use can strengthen the Sustainable Development Goals (SDGs) (G20, 2019).

20.1.3. UNESCO

In November 2019, UNESCO published the Recommendation on the Ethics of Artificial Intelligence addressed to Member States, in their capacity as AI actors and as authorities responsible for the development of legal and regulatory frameworks for AI systems. It also provides ethical guidance to all AI actors, including the public and private sectors, by laying the foundation for an assessment of the ethical impact of AI systems throughout their life cycle (UNESCO, 2019).

20.1.4. European Union (EU)

The EU has also played a prominent role in the regulation of AI, so we cannot fail to mention one of the most recent regulatory instruments that have been presented to regulate it, especially considering that this economic block intends to increase public and private investment over the next decade at least 20 billion euros per year by 2030. Without a doubt, the work that the European Union has undertaken by presenting the proposal for the "Artificial Intelligence Law" (Ziady Hanna, 2023) (CNN Business, 2023) leads the way towards technological regulation.

20.1.5. Meeting of the G7 Digital and Technological Ministers

In April, the ministerial declaration on current and future challenges in the digital society was issued by the G7 (G7, 2023), where the commitment embodied in the Declaration of the Summit for Democracy was reaffirmed, which addresses emerging technologies such as AI, biotechnologies, quantum technologies and points out that they must be shaped in line with democratic principles, highlighting the importance of international debates on interoperability among AI governance frameworks, recognizing that like-mindedness can achieve a common vision between members of this group and hence develop tools for reliable AI, under frameworks and standards that can promote reliability and allow the evaluation of AI.

20.1.6. Spain

A preventive exercise implemented by the Spanish Government is a Sandbox (2022), to guarantee the development of responsible Artificial Intelligence and mitigate the potential risks of this technology for health, safety, and fundamental rights. The purpose is to implement the future obligations of the AI Regulation and other future-proofed supporting materials.

On the other hand, the Council of Ministers agreed a Royal Decree approving the statute of the Spanish Agency for the Supervision of Artificial Intelligence (Ibarra, E., 2023) (AESIA), as a result of the joint work of the Ministry of Finance and Public Function and the Ministry of Economic Affairs and Digital Transformation.

20.1.7. Latin American region

This region has not been left behind in the deployment of instruments focused on issuing guidance and orientation for the proper use of AI, aimed at its governance and regulation. Likewise, regulatory sandboxes have been identified, although they have been developing for several years worldwide, there is still no concrete definition of their scope. However, we could point out that these instruments are possible since there is room for maneuver within

the current regulation that allows for flexible exploration without sanctions for non-compliance with the law.

Some examples of advances in the Latin American region are:

20.1.8. Brazil

In December 2022 (Ferreira, M., 2023) (LexLatin, 2023), the final report of the commission of jurists established to propose a draft regulation of AI in Brazil which was delivered to the Senate. This proposal includes governance measures, with accountability in case of breach of the law, a requirement of transparency in the use of AI, and the guarantee of respect for fundamental rights with the guideline that algorithms do not accentuate forms of discrimination.

20.1.9. Uruguay

Uruguay has been one of the first countries to ratify its intention to implement the UNESCO Recommendation, hence it will also be one of the first countries in the region to apply the Readiness Assessment Methodology (RAM) of UNESCO. Furthermore, currently, the Uruguayan government, through the Agency for Electronic Government and Information and Knowledge Society (AGESIC), is updating its national AI strategy, to propose more relevant and effective actions following the progress in IA (UNESCO, 2023).

20.1.10. Peru

Despite being a country with large digital divide among the population, Peru has dedicated several efforts to promote the advancement of digital transformation in the public sector. In that sense, it published the “Law that promotes the use of Artificial Intelligence (AI) in favor of the economic and social development of the country” (Official Gazette of the Bicentenary “El Peruano”, 2023), one of its first specific steps in favor of this technology.

20.1.11. Chile

Through its AI policy, Chile has decided to advance in regulatory experimentation in the face of the challenges that these systems can generate, allowing this exercise to become a source of knowledge and experience, to motivate innovation in the country and generate conditions for an ambitious deployment of technology in different productive sectors.

The implementation of sandboxes (CAF, 2021) is one of the most ambitious proposals of the AI Policy in Chile and the challenge is to achieve its adequate implementation, promote discussion on the matter, generating a greater understanding of regulatory sandboxes and the impact they are having globally. Chile is also committed to being competitive on a regional and international stage and will adapt the existing regulations of this technology.

20.1.12. Colombia

In Colombia, an instrument was created as recommendations to adopt an ethical framework as a guide for the implementation of artificial intelligence in the national public sector, the objective of this framework (Ethical Framework for Artificial Intelligence in Colombia, 2021), is to fully recognize the need to protect and reinforce all human rights of citizens in the development, use and governance of AI, ensuring their respect and application.

20.1.13. Dominican Republic

The Caribbean region has also shown its interest and according to the Government AI Readiness Index 2022, the Dominican Republic occupies ninth position as one of the most prepared countries in terms of AI. With the support of the Development Bank of Latin America and the Caribbean (CAF) and UNESCO, it has established a Digital Agenda 2030 and a National Innovation Policy 2030, which has established the mandate of developing a National AI Strategy in 2023 (ENIA) (CAF, 2023).

20.1.14. Mexico

Regarding Mexico's efforts, in 2023 the National Artificial Intelligence Alliance (ANIA) was presented and brought together voices of experts and interested authorities to strengthen the AI environment in Mexico through collective consciousness based on human rights, principles of collaboration, and interoperability with an ethical, inclusive, comprehensive, objective, and multidisciplinary perspective.

To assure an inclusive, participatory society that promotes the defense and guarantee of our human rights in the face of technological innovations, the National Institute of Transparency, Access to Information and Personal Data Protection (INAI) will contribute by making personal data protection and privacy an avant-garde framework that serves as a reference in government work and in the private sector, through which the effective protection and guarantee of these rights are privileged (Mendoza, J., 2023).

The Open Loop, Mexico chapter (Del Pozo, et al, 2020), is another exercise coordinated and directed by Meta before Facebook developed a Public Policy Prototype (PPP) focused on transparency and explainability of AI systems to add more value to its users and the general public that will result in a report of public policy recommendations for Mexican regulators.

Some examples of North American advances are:

20.1.15. United States of America (USA)

In USA, some advances have been presented, such as the proposal for the "Algorithmic Accountability Act" (US Senate, 2022)(2022), which addresses the impacts of automated systems and creates new transparency. The "AI Bill of Rights" (The White House, 2022)(2022) was also presented, which is a guide to protect citizens against biases and unequal treatment that data can generate, and the National Institute of Standards and Technology (NIST) published the "AI Risk Management Framework and launched its AI Resource Center" (NIST, 2023)(2023). So far it could be the most detailed framework of its kind in the US.

20.1.16. Canada

In June 2022, the Government of Canada launched the proposal of the "Artificial Intelligence and Data Act (AIDA, 2022)" as part of Bill C-27, the Digital Charter Implementation Act of 2022. The AIDA represents an important milestone in the implementation of the Digital Charter and ensures users trust in the digital technologies they use every day.

21.2. AI Governance

Governments should adopt a regulatory framework that establishes a procedure for public authorities to carry out ethical impact assessments of AI systems to anticipate impacts, mitigate risks, and establish appropriate oversight mechanisms such as auditability, traceability, and explainability, which allow the evaluation of algorithms, data, and design processes (UNESCO, 2019).

Disruptive technologies such as artificial intelligence must be conceived from their design with certain values that respect human rights, democracy, and diversity, since they can amplify inequalities and cause harm, particularly to vulnerable and marginalized groups. Likewise, they must consider appropriate security measures for the benefit of people, through supervision and compliance mechanisms, aligned with principles and rules that allow accountability in any situation.

Our societies must adapt to the transformation that AI will bring through changes in their cooperation framework and governance model. Building a human-centered intelligent society requires the full cooperation of government, businesses, social organizations, and academia. Continuous human monitoring is essential to ensure that algorithms do not lead to unwanted or uncontrolled results.

21.2.1. Principles of Transparency and Accountability

Trustworthy AI depends on accountability, which in turn presupposes that there must be transparency and explainability in the systems related to this technology throughout its life cycle. Transparency reflects the degree to which information about an AI system and its

results are available to people who interact with that system, regardless of whether they are even aware that they are doing so.

Transparency and explainability mean that people know and understand how information is collected and processed and the purposes for doing so, especially when personal data is processed. However, today, there is a so-called black box in the decision-making of AI systems, this means that there are processes in the operation of the system that are not known where they arise from. Therefore, it is essential to join efforts to identify one by one the reasons why the system took a decision since the result can affect the rights, freedoms, and even the security of the individual. Therefore, AI systems must be explainable, the information must be easily accessible and understandable to evaluate their impact, to greater transparency, a better understanding of the functionality of these systems.

To enable accountability, consideration should be given to implementing appropriate monitoring, impact assessment, auditing, and due diligence mechanisms, even concerning the protection of whistleblowers, to ensure accountability across the reporting process throughout its life cycle.

Developers and control authorities are responsible for guaranteeing the auditability and traceability of the functionality of AI systems so that attention and solutions can be given preventively to possible conflicts that may arise concerning possible breaches and threats to human rights.

21.3. Latin American Panorama of AI

Recently, the first Latin American Artificial Intelligence Index (ILIA) was presented, which contemplates an exhaustive analysis of the situation of AI in 12 Latin American countries (Argentina, Bolivia, Brazil, Chile, Colombia, Costa Rica, Ecuador, Mexico, Panama, Paraguay, Peru, and Uruguay). Although this study does not cover all the countries that make up this region, it highlights the efforts that have been made around the implementation of systems that involve the use of AI.

ILIA is a public study that offers a detailed and broad view of the current state of AI in the Latin American and Caribbean region. With a focus on local relevance, the study covers in detail topics such as infrastructure, human capital, data availability, regulations, strategic areas, citizen participation, among others.

One of the main findings revealed by this study is that countries that have a current National AI Strategy show greater institutional insertion and more harmonious regulatory development, as in these topics as in the ILIA in general. Although it is not a common relationship, having a strategy agreed upon at the local stage, seems to be a starting point for other elements regarding the institutional environment.

Likewise, Argentina, Brazil, and Mexico are regional leaders in terms of participation in international spaces to influence the global discussion on AI. Chile shows the best performance in terms of citizen participation in the formulation of strategies. Peru stands out in terms of regulation and legislation related to AI or data protection. However, each country can improve in at least one sub-topic of governance.

The best index scores were obtained by Chile (72.67) and Brazil (65.31), who emerge as the leaders in the region. For its part, Mexico (48.55) has a mature and solid ecosystem, with good performance in research, but there is still a need to strengthen infrastructure, professional training, promote innovation and development.

21.4. Latin American proposal towards the creation of an ex profeso mechanism that assists in matters related to AI.

Due to the analysis of the Latin American Artificial Intelligence Index results, areas and opportunities for improvement have been detected in our region. On one hand, we must guarantee that the existing digital divide decreases as much as possible among nations and populations, besides guarantee equitable access to internet and artificial intelligence systems. On the other hand, it should be considered that it will be the responsibility of the competent authorities in each jurisdiction to set the foundations and recommendations for the appropriate use of this technology, launching training, awareness and dissemination actions for general population, which will allow us to bring a fruitful adoption of a regulation on the matter without affecting or stopping technological innovation, which every day offers more options that make users life easier.

For the above to happen, and to standardize or establish minimum standards for AI regulation over the region, it is proposed to create a Committee of Experts made up of multidisciplinary specialists who will collaborate to investigate, assist, and monitor cases that arise from breaches caused by these technologies against the human rights of personal data protection and privacy of users.

The roadmap ahead to drive this proposal would be under the scheme of the presentation and promotion of resolution CJI/doc. 673/22 rev.1, regarding the draft of inter-American principles on neurosciences, neurotechnologies, and human rights, presented by Dr. Ramiro Orias Arredondo, Member of the Inter-American Juridical Committee and rapporteur on the topic before the Organization of American States (OAS).

Another way to achieve this would be to count on the support of the special rapporteur designated for data protection of the OAS, as occurred with the approval of the Updated Principles on Privacy and Personal Data Protection prepared by the Inter-American Juridical Committee through resolution AG/RES. 2975 (LI-O/21), in November 2021.

This referential regulatory framework will allow us to find the appropriate path so that the proposal has the expected scope and be of interest to the members of the Inter-American

Juridical Committee to position it before the members that make up the General Assembly of this Organization.

As noted, to facilitate the implementation of this proposal, it is desirable to create a Committee of Experts that previously analyzed and agreed on the importance and urgent need to contribute, through non-binding mechanisms, to the situation regarding the use and implementation of existing and yet to be developed disruptive technologies, given the risk they could imply in the private lives of users.

This Committee of Experts must also have powers to direct impact assessments on human rights and ethical impact assessments regarding the use and scope of disruptive technologies. The committee must be made up of multi-stakeholders from different sectors and with multidisciplinary profiles that allow the objective and technical analysis of each of the specific cases that are addressed not only from a privacy perspective but also from a comprehensive perspective.

Once integrated and with a clear proposal on the direction of this initiative, we must open the way for having the support of the bodies within the OAS to present it to the General Assembly and, if considered viable, through a resolution to the Inter-American Legal Committee to integrate the Expert Committee and establish the operating rules for the designation of its members and the cases that would be followed up and attended to.

The Committee would, in principle, be integrated of experts in privacy, personal data protection, and technological developers from both private and public sectors, researchers, representatives of academia, philosophers, sociologists, representatives of regional and international networks, and civil society organizations. The selection procedure would be in two ways, through a public call and through a direct invitation to profiles who, due to their experience, could be part of this select group. Once elected, the content of a resolution and the possible actions to be implemented will be worked on to expand the scope and dissemination of the cases in question and thereby formalize the proposal through the Legal Committee to the rest of the members that make up the General Assembly.

The objective of this Committee of Experts must be grounded on goodwill and its foundations will be on the exchange of knowledge and good practices that promote international cooperation, based on multilateralism and the opportunities that it offers us to strengthen the protection of human rights, joining efforts with other international organizations that have also spoken out on the matter, as well as with groups of economic powerhouses that have shown their concern about this panorama of the new digital age.

In order to guarantee that this proposal is inclusive and considers diverse perspectives, consultative processes will be sought that result in the support and political backing of the parties involved to obtain the desired scope. The work of this Committee would be based on a mechanism that will seek:

- Analyze specific cases

The experts of this mechanism, with the support of the General Assembly, will analyze national laws and institutions and evaluate whether they are effective in preventing breaches of human rights based on the technology in question.

- Issue recommendations

The Committee, mandated by the resolution, will formulate recommendations to States to improve and adapt their legal frameworks and institutions, taking into consideration the principles included in the existing soft law regulations (UNESCO, OECD) in terms of artificial intelligence such as accountability, transparency and explainability.

- Provide follow-up

The Committee will follow up on the recommendations made to each State to evaluate their implementation and the progress made by each country in the matter.

- Develop cooperation tools

Cooperation being as one of its pillars, the Committee will develop tools such as model laws, principles, and legislative guides so that States can count on them when carrying out reforms in their legal frameworks to strengthen them to prevent breaches of human rights, such as privacy and personal data derived from the use of technologies.

References

The White House. AI Bill of Rights, 2022. Available at: https://www.whitehouse.gov/ostp/ai-bill-of-rights/

European Commission. AI regulation sandbox in the EU. 2022. Available at: https://digital-strategy.ec.europa.eu/en/news/first-regulatory-sandbox-artificial-intelligence-presented

NIST. AI Risk Management Framework, 2023. Available at: https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf

US Senate. Algorithmic Accountability Act, 2022. Available at: https://www.wyden.senate.gov/imo/media/doc/Algorithmic%20Accountability%20Act%20of%202022%20Bill%20Text.pdf

Artificial Intelligence and Data Act (AIDA), 2022. Available at: https://ised-isde.canada.ca/site/innovation-better-canada/en/artificial-intelligence-and-data-act-aida-companion-document

COX, Robert W. Multilateralism and World Order. In: COX, R. W.; SASSEN, T. S. (eds.). Approaches to World Order. Cambridge: Cambridge University Press, 1996.

DEL POZO, C. et al. Open Loop Mexico: Public Policy Prototype on the Transparency and Explainability of Artificial Intelligence Systems. 2023. Available at: https://openloop.org/programs/ai-transparency-explainability-mexico/

Development Bank of Latin America (CAF). Impulsando la Inteligencia Artificial en América Latina y el Caribe: Lecciones desde República Dominicana y Uruguay. 2023. Available at: 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/

Development Bank of Latin America (CAF). Sandbox Regulatorio de Inteligencia Artificial en Chile. August 2021. Available at: https://www.economia.gob.cl/wp-content/uploads/2021/09/PaperSandboxIA.pdf

Diario oficial del Bicentenario, El Peruano. Ley que promueve el uso de la inteligencia artificial en favor del desarrollo económico y social del país. July 2023. Available at: https://busquedas.elperuano.pe/dispositivo/NL/2192926-1

FERREIRA, M. Brasil presenta proyecto para regular el uso de la inteligencia artificial. Lexlatin, abril 2023. Available at: https://lexlatin.com/opinion/brasil-regular-uso-inteligencia-artificial

G20. G20 Ministerial Statement on Trade and Digital Economy. 2019. Available at: https://www.mofa.go.jp/files/000486596.pdf

G7. Ministerial Declaration The G7 Digital and Tech Minister's Meeting April 30, 2023. 2023. Available at: http://www.g7.utoronto.ca/ict/2023-ministerial_declaration_dtmm.pdf

IBARRA, E. Aprobado el Estatuto de la Agencia Española de Supervisión de la Inteligencia Artificial. 2023. Available at: 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_%20campaign=share_via

Marco Ético para la Inteligencia Artificial en Colombia. October 2021. Available at: https://minciencias.gov.co/sites/default/files/marco-etico-ia-colombia-2021.pdf

MENDOZA, J. El Compromiso del INAI con la Alianza Nacional de Inteligencia Artificial. The Economist, April 2023. Available at: https://www.eleconomista.com.mx/opinion/El-compromiso-del-Inai-con-la-Alianza-Nacional-de-Inteligencia-Artificial-20230423-0004.html

Organization for Economic Cooperation and Development (OECD). Recommendation of the Council on Artificial Intelligence. 2019. Available at: https://legalinstruments.oecd.org/en/instruments/oecd-legal-0449

United Nations Educational, Scientific and Cultural Organization (UNESCO). Recommendation on the Ethics of Artificial Intelligence. 2019. Available at: https://unesdoc.unesco.org/ark:/48223/pf0000380455

United Nations Educational, Scientific and Cultural Organization (UNESCO). Inteligencia Artificial, Ciudadanía Digital y Educación. June 2023. Available at: https://www.unesco.org/es/articles/inteligencia-artificial-ciudadania-digital-y-educacion

ZIADY, Hanna. Europe is leading the race to regulate AI. Here's what you need to know. CNNBusiness, 2023. Available at: https://edition.cnn.com/2023/06/15/tech/ai-act-europe-key-takeaways/index.html

22. Regulatory Aspects of AI in Argentina

María Julia Giorgelli, Experta independiente.

Abstract

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 carried out by the National Executive Power and summarizes the latest bill projects presented at the national level. In all cases, the focus is on the right to privacy/personal data and transparency/information.

Introduction

In 2022, the Argentine Institute of Statistics and Census reported that 62.6% of urban households have access to a computer and 92.1% to the Internet. Additionally, it announced that 89 out of every 100 people use a cell phone and 88 out of every 100 use the Internet (INDEC, 2022).

Daily, in areas such as work, education, security, justice administration, health, or the exchange of goods and services, various systems based on artificial intelligence are used. It is also employed in less essential sectors, though equally relevant, such as entertainment. As the phenomenon grows rapidly, with a focus on enhancing efficiency and productivity, it also reveals challenges that significantly affect the fundamental rights of individuals.

Various sectors and actors have raised their voices highlighting the problems it entails. Among them, are the numerous criticisms of the French philosopher Eric Sadin (Sadin, E., 2020) or the observations of Kate Crawford (Crawford, K., 2022) who claims that AI is "neither artificial nor intelligent". In the same vein, the Argentine collective work "Thinking digital technology with a gender perspective" argues that these advancements have consolidated the status quo, with situations of inequity and injustice by powerful actors to the detriment of marginalized or minimized sectors (Balmaceda, T., Pedace, K., Lawler, D., 2021) .

We are facing a complex and global product that responds to its time and it is evident that it requires a multidisciplinary approach. Moreover, it should be able to reflect the particularities of each region, provide guarantees to those affected, and establish certain safeguards such as prior testing, human supervision, the obligation to inform and make transparent the processing and use of data, and, if applicable, take responsibility for potential damages.

In summary, work must be done to ensure that AI systems are indeed centered on the public good and not on the market and productivity (Guerra, J., 2023).

22.1. Regulatory situation in Argentina

22.1.1. General Norms

To date, Argentina does not have specific regulations on artificial intelligence. Nevertheless, a normative context relates to it and favours the topic. Particularly, there are pre-existing supranational guidelines and legislation that safeguard offline rights and foster the development of the sector, thereby helping to define and address the issue.

At the supranational level, there are two significant precedents recognized by our country. Both constitute "soft law" and are primarily directed at nation-states; they highlight the complexity of the phenomenon facing humanity and agree on the need for interdisciplinary work, as well as the need for technology to be centred on the common good. These are

principles that will be repeated in various compendiums and constitute the backbone of the issue.

In the "Principles on Artificial Intelligence" signed by thirty-eight countries and eight observers[1], a commitment is established, it aims to achieve robust, safe, impartial, and reliable AI systems. These Principles include a specific provision on "transparency and explainability"[2] aimed at allowing those affected to easily understand the logic that formed the basis for the prediction that harmed them. The protection of personal data is a recurring theme throughout various sections of the guidelines, although it is not treated as an independent concept. Despite this, the need for the development of AI free from biases is emphasized.

The second precedent is the "Recommendation on the Ethics of Artificial Intelligence" issued by UNESCO in 2021. This compendium stands out as the first document of its kind with global reach. On that occasion, the ethical value was considered as the core of the standard, emphasizing the importance of respecting human dignity and protecting the common good. This guideline is denser and broader than the previous one, likely, because it captures the debate of those times. It focuses on aspects related to the environment, diversity, inclusion, and non-discrimination. In these Recommendations, there is a specific section on "protection of the right to privacy and personal data" as well as on "information and transparency"[3]. Regarding personal data and privacy, the standard aligns with a continental-European conception of law, which is of great value for our country that maintains a similar framework. It calls for ensuring the protection of personal information throughout the entire life cycle of the AI system. The need to conduct privacy impact assessments is also mentioned, a tool that will allow, through a preliminary evaluation, to assess the use of such systems. About the information that should be provided to people, it affirms the need to "increase the transparency and applicability of AI systems" and adds the idea that this should be "appropriate to the context".

---

[1] Available at https://www.oecd.org/espanol/noticias/cuarentaydospaisesadoptanlosprincipiosdelaocdesobreinteligenciaartificial.htm

[2] Section 1.3 "AI actors must commit to transparency and responsible disclosure of AI systems. For this, they must provide meaningful information that is context-appropriate and consistent with the state of the art: i. Promote a general understanding of AI systems. ii. Raise awareness among stakeholders about their interactions with AI systems, including in the workplace, iii. enable those affected by an AI system to understand the outcome, and iv. allow those negatively affected by an AI system to question its outcome based on simple and easy-to-understand information about the factors and logic that underpinned the prediction, recommendation, or decision."; available at https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449

[3] Right to privacy and data protection sections 32, 33, and 34. Transparency and Explainability sections 37 and 38, available at https://unesdoc.unesco.org/ark:/48223/pf0000381137_spa

As mentioned, within the domestic legal system, there is legislation that guarantees certain rights and other related laws that identify "technology" as a value to be prioritized and protected.

The National Constitution itself states as an attribute of the National Congress the responsibility to "Provide for human development, for economic progress with social justice, for the productivity of the national economy, for job creation, for professional training of workers, for the defence of the currency value, for research, and for the scientific and technological development, its dissemination, and exploitation." Similarly, the following regulations are identified: "Science, technology, and innovation" (Law No. 25467 of 2001)(Argentina, 2001), the "Regime for the Promotion of the Knowledge Economy" (Law No. 27506 of 2019)(Argentina, 2019), or the most recent on "Financing of the national system of science, technology in innovation" (Law No. 27614 of 2021)(Argentina, 2021).

Particularly noteworthy is the legislation concerning "Personal Data Protection". On one hand, this significance stems from its constitutional recognition and, on the other, from Law No. 25326 of 2000, which offers a specific safeguard intertwining the protection of personal data with transparency as explicit rights. These rights can be asserted against entities processing our personal information (Argentina, 2000, Chapter III, art. 13 and onwards). There is also a specific article on automated data processing, although regarding judicial decisions or administrative acts (art. 20). It is worth noting that this law is in the process of being updated, establishing its application regardless of the techniques, processes, or technologies—current or future—that may arise. Also, the provisions of the "Council of Europe Convention No. 108 for the protection of individuals with regard to the automated processing of personal data and additional protocol to the agreement for the protection of individuals with regard to the automated processing of personal data, the control authorities, and the cross-border flow of data known as Convention 108 (Law No. 27.483 of 2019)" along with its modernized version would apply.

Until December 2023, there was a ministerial-level department called "Science, Technology, and Innovation." Currently, due to the change of government in the presidential elections of December 2023, this department does not exist anymore.

Beyond this, it is important to highlight that in June of last year, the "Recommendations for Reliable Artificial Intelligence" (Argentina, 2023), were issued. It is a guiding, non-mandatory norm. However, it is very interesting since it provides guidelines for public sector innovation projects. It develops a conceptual framework, describes the various AI cycles, and provides recommendations and guidelines for implementing an AI project, thus serving as a practical guide for innovation projects. Moreover, it is the first formal state guideline. Among its core aspects, it guarantees the protection of the right to privacy and data protection along with transparency and explainability from the design stage of the tools, in line with supranational provisions

Also, in 2023, the Administrative Decision 750/2023 (Argentina, 2023) was issued, and with clear coordination purposes of the subject, the Interministerial Table on Artificial Intelligence was created. It aims to address the advancement and application of Artificial Intelligence in various sectors of the economy and society, in accordance with an ethical framework, sustainable development, and digital transformation, and with the purpose of designing an integral strategy on the matter to be applied by the National Executive Power.

The country has an administrative body called the Agency for Access to Public Information, whose purpose is to ensure the fulfilment of the rights to access public information and the protection of personal data. From this area, a "Program for Transparency and Protection of

Personal Data in the Use of Artificial Intelligence" was recently established. Its objectives include: investigating the social, economic, labour, cultural, and environmental implications of AI development in Argentina; analysing the current situation regarding the use of AI in national public sector organizations; generating knowledge to identify good practices, learnings, and recommendations in terms of transparency, algorithmic transparency, and the protection of personal data in the use of AI; carrying out actions to strengthen the institutional capacities of key actors in the implementation of AI in terms of transparency and data protection; and promoting participatory processes for the generation of regulatory proposals for AI in the country, in coordination with other governmental bodies competent in the matter. Currently, despite the change in government, the area has carried out certain actions in the field. Specifically, debates have been developed with experts from the academic and public sectors of different provinces of the country (AAIP, 2024).

Lastly, to complete this scenario, it is interesting to refer to the recent "Latin American Artificial Intelligence Index" (ECLAC., 2023). This is a study carried out on twelve countries of the region with the support of ECLAC and presented in August of this year in Chile. It compares three dimensions: 1. Enabling Factors; 2. Research and Development; 3. Governance. The conclusions are positive regarding Argentina, highlighting a solid foundation for developing and leveraging the potential of AI. It also mentions the need for generating specific regulations in AI, while positively emphasizing the infrastructure in terms of connectivity and a high potential in the development of talent and technological skills. In terms of "governance," the ratings are high.

22.1.2. Legislative Proposals

Below are three legislative proposals considered based on their submissions to the National Congress within the year, hence maintaining parliamentary status[4].

The first project, No. 1472-D-2023[5], proposes the reform of existing law (Law No. 25467 "Science, technology, and innovation"). This proposal is concise, consisting of three articles, and primarily focuses on general aspects based on ethics. It is positive because it anticipates that any advancement in AI must ensure diversity and inclusion, involving the participation of all individuals and groups. It also emphasizes the need to ensure peace and justice. While well-intentioned, the proposal does not match the complexity of the phenomenon as it lacks concrete tools to guarantee these statements and/or address the numerous problems AI can pose, such as potential damages. Noteworthy cases in the country include wrongful detentions through the use of facial recognition systems[6], and inaccurate predictions of teenage pregnancies in the province of Salta[7], where the University of Buenos Aires

[4] The files are 2505-D-2023, 2504-D-2023, and 1472-D-2023. It is also worth mentioning a previous submission in the same National Congress in 2019 aimed at creating a Federal Council of Artificial Intelligence (File 0509-D-2019). There are also some local initiatives, such as the Buenos Aires City project (File 2093-2023) aimed at establishing mandatory training in data and AI. No draft amendments to other laws were analyzed.

[5] Available at https://www.hcdn.gob.ar/proyectos/resultados-buscador.html

[6] As an example, it is cited https://www.pagina12.com.ar/209910-seis-dias-arrestado-por-un-error-del-sistema-de-reconocimien

[7] Consultation material available at https://liaa.dc.uba.ar/es/sobre-la-prediccion-automatica-de-embarazos-adolescentes/

identified "serious technical and conceptual errors, casting doubt on the reported results and compromising the use of such a tool, especially on such a sensitive matter." Finally, the text also lacks mention of personal data protection and transparency.

The other two projects (2504-D-2023 and 2505-D-2023[8]) are autonomous legislative proposals. One aims to create a legal framework for the regulation of AI development and use, and the other to regulate it within the educational field. Both feature a suitable structure, including rationales, objectives, guiding principles, and guarantees. They also incorporate useful tools such as "glossaries" or the need for "training," which help articulate complex topics like AI. Moreover, they feature budget forecasts, vital for the execution of these initiatives.

The "Legal Framework for the Regulation of the Development and Use of Artificial Intelligence" (No. 2505-D-2023) aims to be comprehensive in AI matters. It comprises thirty-three articles, establishes the creation of a specialized AI body, and specifically includes two articles that address the protection of personal data and privacy, as well as ensuring users have the right to comprehend the workings of such systems. Concerning privacy and personal data: "AI systems are required to respect and safeguard the privacy of users and handle their personal data following relevant data protection laws..." (art. 6). On transparency and explainability, it states: "AI systems need to be transparent in their operations, so users understand how decisions are made and results are achieved. It establishes the right of individuals to request explanations of decisions made by AI systems affecting them" (art. 7).

The other legislative proposal, "Law on Regulation and Use of Artificial Intelligence in Education" (No. 2504-D-2023), includes a specific article on privacy and data protection but does not autonomously address transparency. On privacy and data protection, it notes: "Data protection and privacy a) Data collection: Educational institutions and AI providers must obtain informed consent from students or their legal guardians before collecting and using personal data for educational purposes...c) Student rights: Students have the right to access, correct, and delete their personal data, as well as to request the discontinuation of AI use in their education." (art. 4).

Clearly, appropriate regulatory frameworks enhance individuals' rights. However, it's worth questioning whether they alone can prevent biases or enable us to counteract negative or erroneous predictions. Crawford reflects on industry practices and, when examining facial recognition systems, asserts that images "are not seen as individuals, but as part of a shared technical resource; they are just another data component of the facial recognition verification test program, the quintessential reference in this field." She further notes, "Databases herald the emergence of a logic that has now invaded the tech sector: the belief that anything can be data and that data is there for the taking by whoever wants it. It doesn't matter where the photo was taken, whether it reflects a moment of vulnerability or pain, or if it represents a form of humiliation for the subject. The normalization of taking and using whatever is available across the entire industry means few pauses to question the policies underlying these actions." (Crawford, K., 2022).

In summary, while explicit legal provisions will undoubtedly support individuals in asserting their rights, this alone is not sufficient. It must also be accompanied by a range of actions such as providing resources to competent authorities, conducting outreach or awareness

[8] Available at https://www.hcdn.gob.ar/proyectos/resultados-buscador.html y https://www.hcdn.gob.ar/proyectos/resultados-buscador.html

campaigns, maintaining proactive public policies in favour of individuals, and implementing enforcement actions to realize changes and course corrections.

22.2. Conclusions

There is much to discuss regarding the AI phenomenon.

We recognize it as a tool impacting various fields such as social relationships, subjectivity, law, the environment, and the economy. Despite acknowledging its positive advancements, it's evident that various challenges remain, such as preventing the perpetuation of biases and addressing various harms and infringements on fundamental rights.

These responsibilities should not rest solely with the industry. To believe that this sector will prioritize the common good and the protection of individuals is an illusion. As French philosopher Eric Sadin suggests, these systems embody a colonialist logic (Sadin, E., 2020), which is why it is crucial to work towards representing the particularities of countries in the global South.

Therefore, it is necessary to urgently work in an interdisciplinary manner to achieve an AI focused on the common good.

References

BALMACEDA, T.; PEDACE, K.; LAWLER, D. Thinking about technology from a gender perspective. 2021. Available at: https://proyectoguia.lat/wp-content/uploads/2022/06/perspectiva-generoV6.pdf.

Chief of Cabinet of Ministers, Argentina. Provision 2 / 2023 Recommendations for Reliable Artificial Intelligence. 2023. Available at: https://www.argentina.gob.ar/sites/default/files/2023/06/recomendaciones_para_una_inteligencia_artificial_fiable.pdf.

CRAWFORD, K. Atlas of Artificial Intelligence: Power, Politics, and the Planetary Costs. Fund of Economic Culture, 2022.

AAIP. Córdoba, Rosario, Salta y Santiago del Estero reflexionaron junto a la AAIP sobre Inteligencia Artificial y Protección de Datos Personales. Argentina.gob.ar, February 9, 2024. https://www.argentina.gob.ar/noticias/cordoba-rosario-salta-y-santiago-del-estero-reflexionaron-junto-la-aaip-sobre-inteligencia.

Argentina. Ley 25.467. InfoLeg, September 2001. Available at: https://servicios.infoleg.gob.ar/infolegInternet/anexos/65000-69999/69045/texact.htm.

Argentina. Ley 27506. InfoLeg, June 2019. Available at: http://servicios.infoleg.gob.ar/infolegInternet/anexos/320000-324999/324101/texact.htm.

Argentina. Ley 27614. InfoLeg, 2021. Available at: http://servicios.infoleg.gob.ar/infolegInternet/anexos/345000-349999/347804/norma.htm.

Argentina. Ley 25.326. InfoLeg, October 2000. Available at: https://servicios.infoleg.gob.ar/infolegInternet/anexos/60000-64999/64790/texact.htm.

Argentina. Disposición 2/2023. Boletín Oficial, June 2023. Available at: https://www.boletinoficial.gob.ar/detalleAviso/primera/287679/20230602.

Argentina. Decisión Administrativa 750/2023. Boletín Oficial, September 2023. Available at: https://www.boletinoficial.gob.ar/detalleAviso/primera/293710/20230908.

ECLAC. Latin American Artificial Intelligence Index. 2023. Available at: https://indicelatam.cl/wp-content/uploads/2023/08/CAP-G-ARGENTINA.pdf.

GUERRA, J. Towards a Feminist Framework for AI Development: From Principles to Practice. Latin America: Digital Creative Commons Attribution 4.0 International, 2023.

INDEC. Permanent Household Survey (EPH) Access and Use of Information and Communication Technologies. Fourth quarter of 2022. Available at: https://www.indec.gob.ar/uploads/informesdeprensa/mautic_05_239BB78E7691.pdf.

SADIN, E. ARTIFICIAL INTELLIGENCE OR THE CHALLENGE OF THE CENTURY. Caja Negra, 2020.

23. AI and neurotechnologies: the need for protection in the face of new crossroads

Natalia L. Monti, Fundación Kamanau, Abogada, Magíster en DDHH, integra el Centro de Protección de Datos Personales de la Defensoría del Pueblo CABA, Argentina.

To date, there is no normative text that reviews the applied incidence that science can have on the physical and psychological integrity of human beings and how it could affect their right to life and their physical and psychological integrity, and there is a need for protection in the face of these new crossroads (C. S. Girardi c/ Emotiv Inc, 2023).

Abstract

With regard to the OAS Declaration of Inter-American Principles on Neurosciences, Neurotechnologies and Human Rights, approved in March 2023, and in addition to the recent ruling of the Chilean Supreme Court against the company that manufactures and markets neurotechnology devices, Guido Girardi v. Emotiv, of August 2023, it is necessary to establish regulatory criteria that clarify international standards on new advances in science and technology, always from a human rights perspective. There is a need to establish regulatory criteria that make clear international standards on new advances in science and technology, always with a human rights-based perspective. The challenges will be to generate reliable scenarios in the development of these new technologies, especially immersive technologies that use neurotechnology and artificial intelligence for the most vulnerable sectors, such as children and adolescents, genders and dissidences, and persons with disabilities, among others.

23.1. Developments in the scientific field

In recent years, the accelerated development of new artificial intelligence (AI) technologies is raising many concerns, especially in relation to the ethical and legal criteria on which these innovations in science should be based.

Thus, in relation to the protection of human rights, different initiatives are being promoted, due to the growing impact of these new scientific advances not only on the life of societies, but also because of the effect they can have on freedom, thought and physical integrity, generating a new frontier, previously little known to the legal world.

In this sense, it is known that certain applications allow a bidirectional connection between an individual's central nervous system and an electronic system. This is how the possibility of accessing the data produced by brain information and exploring them, recording them in external devices, as well as deleting and even modifying them appears (Yuste Rafael, G. J., 2021).

At the heart of neurotechnology are *Brain Computer Interfaces* (BCIs), which are devices that connect the human brain to a computer or other device outside the human body. In this sense, neurotechnological devices (Yuste, R., 2023) can be implanted or non-invasive elements (glasses, helmet, headband, bracelet, etc.) that can be electronic or electrical, can be chips, optical, magnetic, acoustic, molecular or chemical, and increasingly use AI.

Thus, we see how modern advances in neuroscience and neurotechnologies have unlocked the human brain and allowed us to learn more about brain processes and their relationship to mental states and observable behaviour (Ilenca, M. A., 2017).

It is clear that the development of neurotechnologies can have a positive impact on people's quality of life and health. In that sense, neurotechnologies offer enormous potential in the medical field for the treatment of neurological and mental disorders. There are more than

300,000 different mobile health applications available worldwide (a number that has doubled in just five years), with an estimated market value of more than $100 billion (“The rise of...”, 2019). Consequently, it is intended to improve our scientific understanding of human brain function and unlock the pathological riddles of various treatment-resistant neurological and mental disorders (UNESCO, C. I., 2021).

At the same time, such developments are increasingly being applied in contexts outside of healthcare, entering fields such as education, the workplace and entertainment, among others. Globally, the neurotechnology market is growing at a compound annual growth rate of 12 percent and is expected to reach $21 billion by 2026 (Expert Market Research , n.d.).

In this sense, there are non-invasive devices that can very simply, for example, evaluate how players feel when they are exposed to different stimuli and, based on this, different actions can be tried out, with the aim of boosting performance. Likewise, "it is also being used in universities to study cognitive development in children, particularly in disadvantaged populations", as the company Emotiv explained some time ago (Jaimovich, D., 2017).

In this way, measures can be taken to improve attention, reduce stress or improve concentration. A range of available devices can be purchased from the Emotiv website[1] , and anyone interested can obtain them for as little as US$499. The device (headband) collects information and allows an electroencephalogram (EEG) to be taken in just seconds and wirelessly, stored and can be used to analyse the impact of different external factors on the person and their emotions (Jaimovich, D., 2017).

Understanding the importance of the human brain in its mental and cognitive functions, it is necessary to question ourselves as a society to what extent we consider interference with brain activity legitimate, whether there are current regulatory limits to its implementation, or whether we need to deepen regulations that specify clear criteria for these developments.

23.2. Regulatory initiatives to minimise human rights impacts

In recent times, various initiatives have been developed to minimise the impact of neurotechnologies in relation to their application in AI. The focus is always on the protection of human rights, but there are also ethical issues at stake.

That is to say, it is important to highlight that while there are several technologies that pose risks to human rights, we will refer specifically to AI with the use of neurotechnology.

In this sense we understand AI as the "constellation of processes and technologies that enable computers to complement or replace specific tasks that would otherwise be performed by humans, such as making a decision or solving a problem" (UN, 2021).

On the other hand, AI could be misused and provide tools for manipulation, exploitation and social control (European Parliament and European Council, 2021). Therefore, these risks could have a greater impact on people who are in a special situation of vulnerability, such as children and adolescents, gender and dissidence, people with disabilities, ethnic and racial minorities, the elderly, people living in poverty, among others.

[1] https://www.emotiv.com/

At the international level, several organisations have begun to work on the issue. Notably, in its 2019 "Recommendation on responsible innovation in neurotechnology", the OECD[2] mentions the need for safeguards for mental information.

Likewise, in 2020, the draft of the International Bioethics Committee of the United Nations Educational, Scientific and Cultural Organisation ("UNESCO") on *"Ethical Issues in Neurotechnology"* was published, which in its conclusions affirmed the need to provide a framework for the development of neurotechnology from a human rights perspective, advancing in some conceptual definitions[3]. In December 2021, the report of the UNESCO International Bioethics Committee on Ethical Issues and Neurotechnology was published (UNESCO, 2022). Finally, UNESCO convened a meeting in July 2023 at its headquarters in Paris, at which the possibility of generating a global regulatory framework for neurotechnologies was discussed, similar to that established by the Universal Declaration on the Human Genome in 1997 (UNESCO, 2023) .

On the other hand, an important step forward was made in October 2022, when the United Nations Human Rights Council approved by consensus resolution A/HRC/51/L.3 on "Neurotechnologies and Human Rights". This initiated a study on the impacts, opportunities and challenges of neurotechnologies and generated a consultative process with state actors, multilaterals, the private sector and civil society[4].

Furthermore, in February 2020, the European Commission published the "White Paper on AI: A European approach to excellence and trust", which identified options to reconcile AI developments with mitigating the risks of certain AI uses (Comisión Europea, 2020) which defined the options for making AI developments compatible with mitigating the risks of certain uses of these technologies. This proposal also attracted interest among actors linked to neurotechnologies, because of the deep connection between neurotechnologies and AI. In particular, its contents on the risks of discrimination and the references to the behavioural alterations that they can induce attracted attention.

23.2.1. Inter-American Principles on Neurosciences, Neurotechnologies and Human Rights

At the Inter-American regional level, significant progress has been made with the development of international standards through the work of the Inter-American Juridical Committee (CJI) of the OAS.

In this sense, in a novel way, the IJC approved the "Declaration on Neuroscience, Neurotechnologies and Human Rights: New Legal Challenges for the Americas" (2021), which was the first of its kind worldwide (CJI, 2021), in one of its sections the Declaration makes it clear that there are no specific regulations, which is why it is essential to call on the actors to pay attention and safeguard the human rights of individuals in the face of the dizzying technological development.

---

[2] Available at https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0457

[3] Available at https://is.gd/my7DCF. On the other hand, we highlight the impulse of the Bioethics Committee of the Council of Europe, which published a Strategic Action Plan on Human Rights and Technology in Biomedicine (Council of Europe, n.d.).

[4] Among the antecedents that motivated this initiative, the advances at the Ibero-American level and the Chilean constitutional reform on the protection of brain activity and information were mentioned (UN, 2022).

In this sense, the Declaration warns that the advances in neuroscience and the development of neurotechnologies require deep reflection by all sectors involved, calls on States, the private sector, academia and the scientific world, and requests the adoption of concrete measures by each of the actors that will allow these innovations to contribute to the common good.

Subsequently, the OAS Inter-American Juridical Committee continued its work to develop more precise standards to help guide and harmonise the necessary national regulations in this area.

Since then, various actions have been carried out[5], including preparatory work with the Committee of Experts that had collaborated in the drafting of the Declaration[6] , to deepen the work and generate minimum principles to recommend to States the treatment of neurotechnologies. Thus, the IAJC approved a second progress report entitled "Draft Inter-American Principles on Neuroscience, Neurotechnologies and Human Rights" (OAS, 2022).

Finally arriving on 9 March 2023, when the IJC approved the document on **Inter-American Principles on Neurosciences, Neurotechnologies and Human Rights**[7].

The development of these Principles is the result of an analysis of the international norms and standards that are already in force and are enforceable for States. In this case, greater precision is required on the specific subject of the development of neurotechnologies, in order to anticipate and combat any situation that tends to violate the human rights of individuals.

The document establishes ten sections that reinforce the existing guarantees for the protection of human rights in this area, with the fundamental premise being the preservation of individual identity and cognitive integrity in the face of any neurotechnological intervention.

In addition, certain standards derive from an in-depth interpretation of personal data protection principles. Thus, it establishes that the rights of individuals should be protected from the beginning of the design of neurotechnologies until their final deployment, evaluation, commercialisation and use. On the other hand, it seeks to provide greater protection for neural data, declaring them to be sensitive data. It also establishes the importance of having express consent to undergo any device that involves the manipulation of neurodata.

On the other hand, in relation to the protection of the right to equality, it provides for equal access to neurotechnologies and guarantees non-discrimination of categories that have

---

[5] On 21 June 2022, a public hearing was held before the Inter-American Commission on Human Rights, at the request of a group of experts whose main goal was to identify the challenges, impacts, risks and possible violations of human rights that the unregulated use of neurotechnologies generates - or could generate - in the field of people's health, both on the part of States and the private sector. Likewise, the aim was to show and share recommendations in order not to incur in conduct that is risky for Human Rights. This request for a hearing was made by the Neuro Rights Initiative of Columbia University, the Kamanau Foundation, the Pro Bono Network of the Americas, and the Ronda Foundation, together with a group of experts of different nationalities and professions linked to the world of science and international human rights law, who participated in the preparation of a document for the Americas that would address the challenges and impacts of neurotechnology on human rights and, in particular, on neuro-rights. 184th session of the IACHR: https://www.youtube.com/watch?v=-JdUHdlXgdE

[6] The Committee of Experts is composed of: Eduardo Bertoni, Ciro Colombara, Francesca Fanucci, Verónica Hinestroza, Amelie Kim Cheang, Tomás Quadra Salcedo, Moisés Sánchez, Silvia Serrano Guzmán and Rafael Yuste.

[7] OAS (2023).

historically been subject to discrimination: race, colour, gender, nationality, religion, social status, among others. On the same argument, the need to establish clear limits and to exercise reinforced control over the increase of cognitive capacities is also raised.

Finally, strategies for the efficient governance of neurotechnologies are promoted, establishing oversight and supervisory bodies and ensuring access to effective guardianship.

The Inter-American Principles provide a basis of standards on which States can take action in the face of developments in the scientific world.

We highlight that a few days ago, in August 2023, the IAJC approved a new agenda to delve deeper into this topic, addressing the impact of AI-based technologies on human rights, with a special focus on children and adolescents, based on neurotechnologies, immersive and emerging technologies with application or based on AI.

23.3. Recent Chilean Supreme Court ruling on neurotechnologies

On 9 August 2023, the Chilean Constitutional Court ruled in favour of a complaint regarding the sale and marketing in Chile of Emotiv's Insight device, on the grounds that this device does not adequately protect the privacy of its users' brain information, infringing the right to mental integrity, physical and psychological integrity and the right to privacy.

Let us recall that on 14 October 2021, Chile enacted Law No. 21.383, which amended the Political Constitution of the Republic to establish that scientific and technological development will be at the service of people and will be carried out with respect for life and physical and mental integrity. The constitution states that the law shall regulate the requirements, conditions and restrictions for its use on people, and shall **especially protect brain activity, as well as the information derived from it** (art. 19, inc. 1°). In this sense, the Constitutional Court understood that it constitutes a direct mandate of protection, in addition to various international instruments that recognise the relationship between science and human rights (C. S. Girardi v. Emotiv Inc., 2023).

However, we will analyse separately some issues that are worth highlighting:

23.3.1. Emotiv Insight Device (neurotechnology)

In line with the scientific context outlined above, the bioinformatics company Emotiv develops portable electroencephalography (hereafter "EEG") products[8], including neuro-earpieces, software development kits (hereafter "SDKs"), software, mobile applications and data products. The company is currently headquartered in San Francisco, USA. According to the company, its mission is: "to empower people to understand their own brains and accelerate brain research worldwide".

Currently, Emotiv stands out mainly for the design of two devices: Emotiv Insight and EPOCx which are EEG devices that through non-invasive neuroimaging techniques of functional exploration of the central nervous system, obtain the recording of a person's electrical brain activity in real time. EEG measures the electrical activity of the brain in a very simple way, through the placement of electrodes on the surface of the scalp.

[8] Electroencephalography (EGG) is the recording and evaluation of electrical potentials generated by the brain and obtained by means of electrodes placed on the surface of the scalp. Originally used in the fields of psychology, medicine and neuroscience, it is now widely used in human-computer interaction, gaming, neuromarketing, simulations and others.

The Insight device is positioned as a non-invasive Brain-Computer Interface (BCI) type device. It is wireless and through a headband covers the frontal, temporal and parietooccipital locations around the brain. Designed for everyday use using hydrophilic polymer sensors, this device allows the user to read their emotions and move items - both digital and real - with their mind.

As the company itself puts it, this device provides access to EEG data, consisting of electrical bio-signals that include information about the user's gestures, movements, preferences, reaction times and cognitive activity (Emotiv, n.d.).

However, while current technology does not yet allow us to read thoughts, **neuroimaging techniques have the capacity to record brain activity. An individual's mental substratum is a product of his or her brain activity. The protection of this innermost self, of the subject's inner subjective experience, forms a unique individual sphere whose protection is inseparably linked to the protection of his or her human dignity**.

23.3.2. (Highly sensitive) brain data in the enterprise cloud

In this regard and due to Mr. Guido Girardi Lavin's interest in neurotechnology devices and his concern regarding the risks that may comprise the privacy of brain information, on 28 February 2022, he purchased the Insight device through the Emotiv website. After paying for shipping to UPS, on 21 March 2022 the Insight device arrived at his home address. Following the instructions on the device and in order to record and access his brain data, on 7 April 2022, he created an account on the Emotiv data cloud, called Emotiv Cloud. At that time, Emotiv asked to accept the company's terms and conditions. When Mr Girardi attempted to start recording his brain data, Emotiv advised that because he used the free licence and not the PRO licence, he could not export or import any records of the brain data. In fact, Emotiv pointed out that the data would be held in Emotiv's cloud until Emotiv purchased the Pro licence. Minutes later, the system alerted him that the recorded brain data had been successfully uploaded to the Emotiv cloud.

The Court of Appeal considered that neurodata consists of all the information related to brain activity obtained through the use of advanced neurotechnologies. Neurodata is part of the internet of bodies, the advance of AI brings us to a horizon very close to superintelligence or second-level AI that no longer needs human intervention[9].

The question here is**, does Emotiv adequately protect the privacy of users' brain information, especially with regard to highly sensitive information such as neurodata?**

The company's response was to rely on the pseudonymisation of stored data and on the other hand warned that no security measures are 100% effective and that they cannot guarantee the security of users' personal information.

23.3.3. Consent and purpose (unclear)

It is also noted that another of the arguments in the lawsuit focused on alleging that the use of the device and the storage of their brain data in the Emotiv company exposes users to the risk that it will be shared with third parties and that such data will be used for scientific research and statistical information for free use.

[9] Nevertheless, and under the criterion of users' right to informational self-determination, the claim was dismissed and later reached the Supreme Court. (C.A. Girardi v. Emotiv Inc., 2023).

So, **is the consent given through the platform sufficient for the company Emotiv to store neurodata, and on the other hand, is it sufficient that within the terms and conditions it is stated that this data can be shared with third parties for very generic purposes?**

The company's response was to state that users must expressly consent to the processing of personal and brain data.

The Supreme Court referred to the Chilean Law 20.120 which develops a broad article on the expression of consent to scientific research on human beings (art. 11).

In this sense, the law states that informed consent exists when the person who must give it is aware of the essential aspects of the research, especially its purpose, benefits and risks. The law also clarifies that adequate, sufficient and comprehensible information must have been provided. On the other hand, special mention must be made of the right not to authorise the research or to revoke consent at any time.

For all these reasons, the Supreme Court considered that "the explanation of the requested party, to the effect that the data it obtains from Insight users, on being anonymised, become freely usable statistical information, omits as a prior question the need for express consent for its use for scientific research purposes, other than statistical registration, and expressly regulated in Chile" (C. S. Girardi v. Emotiv Inc., 2023).

The Supreme Court continued that it would rule out the possibility that such consent could be considered as tacitly given through other consents, by those who as consumers acquire certain devices and should have been required to give specific consent indicating the purpose and aim of a certain research with their neurodata.

23.3.4. The active role of the state and the precautionary principle?

In its ruling, the Chilean Supreme Court considered that with the development of new technologies that involve more and more aspects of the human person, which were unthinkable a few years ago, special attention and care must be given to their review by the State.

It is worth noting that this point of the **judgment reflects the precautionary principle brought from environmental law, which is exercised in the face of a potential risk, the action of the State through precautionary measures.**

In this sense, the precautionary principle has been understood for several years now as an attitude that must be observed by those who make decisions concerning an activity that may reasonably be risky for the health or safety of present or future generations (Kamada, L., 2012).

This, in order to prevent and anticipate its possible effects, as well as to directly protect human integrity, which includes privacy, confidentiality and the rights of the psychic integrity and the subject of scientific experimentation (C. S. Girardi v. Emotiv Inc., 2023).

Thus, the Supreme Court considered that at present, regarding the use of these technologies "it is absolutely necessary that prior to allowing their commercialisation and use in the country, this technology and devices be analysed by the relevant authority, understanding that it raises problems that have not been studied before" (C. S. Girardi v. Emotiv Inc., 2023).

Therefore, the Supreme Court stated that the prior assessment of "the handling of the data obtained from it (device) must be strictly in accordance with the applicable regulations" (C. S. Girardi v. Emotiv Inc., 2023) .for the purposes of marketing and use of the Insight device, should be made by the Health Authority, which in this case is the application authority for authorising products or elements for medical use of the Ministry of Health, and the Customs

Authority, so that it can evaluate whether to grant the corresponding customs clearance certificate.

23.3.5. Mental privacy (neuro-rights)

In addition, to conclude with the final part of the ruling, the Supreme Court observed that the conducts carried out by Emotiv violated the constitutional guarantees contained in art. 1 (physical-psychic integrity-brain activity) and 4 (protection of personal data) of article 19 of the Constitution.

This, **contemplating the scope of the constitutional reform that grants greater protection and protection to the development and use of technologies that access and/or modify brain information without respecting the right to life and the physical and psychological integrity of people**. In this sense, "it was ordered to eliminate all the information that had been stored in its cloud or portals, in relation to the use of the device" (C. S. Girardi v. Emotiv Inc, 2023) .

23.4. Need for protection: do we regulate? How?

Taking into account the OAS Inter-American Principles, the initiatives being developed at the international and national level and recent Chilean jurisprudence, there are great challenges ahead.

In this sense, we are aware of the need to establish regulatory criteria that make clear international standards on new advances in science and technology, always with a human rights-based perspective.

In conclusion, and in a very brief way, we will leave some questions that we understand to be key to outline regulations in terms of neurotechnologies, based on AI, or with application in it.

- Do data protection laws need to be strengthened by clarifying that neural data are sensitive personal data? That they acquire maximum protection in terms of disclosure, security and transfer.
- Is the precautionary principle being brought to bear on this issue?
- It is important that express and specific consent is established and the purpose is adequately detailed Is pseudonymisation sufficient to protect privacy?
- Is there a need to implement a compliance model called "accountability" or "proactive responsibility" for risk governance?
- Enforcement authority: what is the ideal institutional model, and is the health and personal data authority sufficient to take control of AI and neurotechnologies?

It is clear that neurodata is bursting onto the legal scene, raising many questions. They will bring together the criteria established by recent regional standards on neurotechnologies regarding the limits within which States would guarantee the protection of human rights. In addition, we believe that the arguments assessed by the Chilean Supreme Court's ruling should be considered in the specific case. All the reasoning should be taken into consideration in future regulations.

In this sense, this new way of living (Farahany, N. A., 2023) is putting under much more intense pressure the regulatory infrastructure needed to allow and sustain all these developments to happen without infringing on human rights. It is clear that developments in neurotechnologies and AI require regulatory clarifications from states, which - in turn - need to rely on international human rights treaties and existing national norms.

We have great challenges ahead of us to generate reliable scenarios for the development of these new technologies, especially immersive technologies using neurotechnology and AI.

It is urgent that clear rules are established and that the rights of the most vulnerable people are particularly protected, such as children and adolescents, people with disabilities, women and dissidents, and older people.

References

184th session of the IACHR. Available at: https://www.youtube.com/watch?v=-JdUHdlXgdE

UN. Draft Resolution A/HRC/51/L.3 meeting files. 2022. Available at: https://hrcmeetings.ohchr.org/HRCSessions/RegularSessions/51/DL_Resolutions/Forms/ResolutionDS/docsethomepage.aspx?ID=12&FolderCTID=0x0120D520005A4381ABFFD48642897E02288D058A22001B07C878276D3B4E9F9B79D83234987E&List=f97dc3a9-0289-4520-a186-336c6365e37d&RootFolder=%2FHRCSessions%2FRegularSessions%2F51%2FDL%5FResolutions%2FA%5FHRC%5F51%5FL%2E3&RecSrc=%2FHRCSessions%2FRegularSessions%2F51%2FDL%5FResolutions%2FA%5FHRC%5F51%5FL%2E3.

C. S. Girardi c/ Emotiv Inc., 105065-2023 (Tercera Sala agosto 9, 2023)

C. S. Girardi v. Emotiv Inc., 105065-2023 (Third Chamber August 9, 2023)

C. S. Girardi v. Emotiv Inc., 105065-2023 (Third Chamber August 9, 2023); C.A. Girardi v. Emotiv Inc., 49852-2022 (Court of Appeals 24 May, 2023)

Comisión Europea. LIBRO BLANCO sobre la inteligencia artificial - un enfoque europeo orientado a la excelencia y la confianza. February 19, 2020, available at: https://eur-lex.europa.eu/legal-content/ES/TXT/PDF/?uri=CELEX:52020DC0065.

OAS. Draft Inter-American Principles on Neurosciences, Neurotechnologies and Human Rights. August 25, 2022. Available at: https://www.oas.org/es/sla/cji/docs/CJI-doc_673-22_rev1_ESP.pdf

OAS. CJI/RES. 281 (CII-O/23) corr.1. March 9, 2023. Available at: https://www.oas.org/es/sla/cji/docs/CJI-RES_281_CII-O-23_corr1_ESP.pdf.

European Parliament and European Council. "Regulation laying down harmonised rules in the field of AI (Artificial Intelligence Act)", 2021, p. 24, par. 15

Expert Market Research. Global Neurotechnology Market Report and Forecast 2022-2027, Report Summary. Available at: https://www.expertmarketresearch.com/reports/neurotechnology-market

Emotiv. Privacy Policy. Available at: https://id.emotivcloud.com/eoidc/privacy/privacy_policy/.

FARAHANY, N. A. The Battle for your Brain. New York: St. Martin's Publishing Group, 2023

ILENCA, M. A. "Towards new human rights in the age of neuroscience and neurotechnology". Life Science and Policy, 13:5, p. 2, 2017

JAIMOVICH, D. How Emotiv Insight works, the headband to move objects with the mind and control emotions. INFOBAE, 9 fevereiro 2017. Available at: https://www.infobae.com/tecno/2017/02/09/asi-funciona-emotiv-insight-la-vincha-para-mover-objetos-con-la-mente-y-controlar-las-emociones/

Comité Jurídico Interamericano (CJI). CJI/DEC. 01 (XCIX-O/21). August 11, 2021. Available at: http://www.oas.org/es/sla/cji/docs/CJI-DEC_01_XCIX-O-21.pdf

KAMADA, L. From the paradigm of certainty to the paradigm of uncertainty as a criterion for judicial decisions in environmental matters. SAIJ, Id SAIJ: DACF120104, 2012

Council of Europe. Strategic Action Plan on Human Rights and Technologies in Biomedicine (2020-2025). Available at: https://rm.coe.int/strategic-action-plan-final-e/1680a2c5d2

The Rise of mHealth Apps: A Market Snapshot, Best Practices (blog). Liquid State, 26 março 2018, atualizado em 12 novembro 2019. Available at: https://liquid-state.com/the-rise-of-mhealth-apps-a-market-snapshot/

UN, Report of the High Commissioner for Human Rights, Michelle Bachelet. "The right to privacy in the digital age", p. 2, 2021

UNESCO, C. I. Ethical issues in neurotechnology. UNESCO. Adotado pelo International Bioethics Committee na 28ª sessão em dezembro, 2021. Available at: https://unesdoc.unesco.org/ark:/48223/pf0000378724

UNESCO. "There is an urgent need to establish an ethical framework on neurotechnology at international level", 6 junho 2023. Available at: https://news.un.org/es/story/2023/06/1521747

UNESCO. Report of the International Bioethics Committee of UNESCO (IBC) on the ethical issues of neurotechnology. UNESDOC, 2022. Available at: https://unesdoc.unesco.org/ark:/48223/pf0000378724.

YUSTE, R. 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. Available at: https://www.youtube.com/watch?v=9Yod5FcNwMo

YUSTE, R.; G. J. "It's time for neuro-rights". Horizon Magazine, N° 8, p. 154-156, 2021

Pre-print version of Belli L. Gaspar. W.B. The Quest for AI Sovereignty, Transparency and Accountability. Springer-Nature. (2025).

24. Conclusion: Harnessing Multistakeholder Governance for Advancing AI Sovereignty, Transparency, and Accountability

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 paper offers a conclusion for the volume's core task of exploring AI sovereignty, transparency, and accountability. These debates are far from settled and an expanding body of literature, from a wide range of stakeholders is contributing to enlighten the many challenges and opportunities inherent in AI governance. Hence, multistakeholder cooperation emerges as indispensable in confronting these challenges and identifying these opportunities. The chapter argues that multistakeholder cooperation, both in policy development and implementation, emerges as a fundamental imperative for achieving responsible and inclusive AI governance that aligns with societal values and aspirations. However, it also emphasises that stakeholders engaging in AI governance face significant challenges both as regards achieving a shared understanding of AI sovereignty, transparency, and accountability, and the design and implementation of strategies and policies able to realise such concepts in an effective way. To conclude this chapter calls for the elaboration of a Multistakeholder Framework for AI Sovereignty, Transparency, and Accountability, an idea that could be further explored by the Data and AI Governance Coalition and UN IGF stakeholders.

Introduction

The objective of this volume has been to embark upon the challenging task of exploring AI sovereignty, transparency, and accountability. As demonstrated by the various contributions to this book, these debates are far from settled and an expanding body of literature, from a wide range of stakeholders it contributing to enlighten the many challenges and opportunities inherent in AI governance. Hence, multistakeholder cooperation emerges as indispensable in confronting these challenges and identifying these opportunities.

Throughout this book, we have scrutinized the fundamental rationales of AI sovereignty and how the elements underpinning this novel concept de facto apply of the concrete case studies of Brazil, South Africa, and India. Importantly, we have emphasised the key role that AI sovereignty discussions will play for nations in the Global South, as they are grappling to understand, develop and regulate AI systems.

In this context, AI transparency and accountability play an instrumental role to make sure that individuals, enterprises, governments, and societies at large can understand the functioning of AI systems, be able to develop them rather than being mere consumers of them, and ultimately, regulate them according to their values. In a nutshell, AI transparency and accountability are vital for AI to be sovereign. However, as we have demonstrated neither AI transparency nor AI accountability are universally defined concepts, thus adding a further layer of complexity to the task of having a meaningful regulation and sound governance of these essential principles.

As we look to the future of AI governance in general and, particularly, to the possible progresses of the AI sovereignty, transparency, and accountability areas, it is imperative that stakeholders collaborate, innovate, and adapt in response to the evolving challenges and opportunities brought by AI. Critically, the contributions to this volume converge in arguing that by embracing a human-centric approach to AI governance, grounded in principles of openness and inclusivity, we can unlock the full potential of AI to advance human welfare, promote sustainable development, and build a more equitable development, while promoting good AI sovereignty[1].

However, we must temper our enthusiasm with a grain of pragmatism, as numerous challenges remain ahead of us. In this concluding chapter we try to address some of these challenges, pointing out some path for future research and multistakeholder debate.

24.1. Key Challenges in AI Sovereignty, Transparency, and Accountability

Stakeholders engaging in AI governance face significant challenges both as regards achieving a shared understanding of AI sovereignty, transparency, and accountability, and the design and implementation of strategies and policies able to realise such concepts in an effective way. Each of them presents unique complexities that must be addressed to foster responsible and sustainable AI development and deployment. Three macro-categories of challenges stand out particularly prominently: promoting a positive or "good"[2] vision of AI sovereignty, resisting to authoritarian versions of sovereignty; overcoming the conceptual ambiguity of transparency, making the concept meaningfully applicable; and addressing the oversight and governance gaps of accountability mechanisms, too often considered as a secondary preoccupation. Let us unpack these challenges briefly.

First, while we may well argue that AI sovereignty is essential for safeguarding national interests and strategic autonomy, we must acknowledge that sovereignty narratives can also be exploited to implement trade restrictions or embed authoritarian features into AI systems[3]. These risks are concrete and do not only affect domestic affairs: in the current digital cold war scenario, there is a growing menace of weaponizing AI supply chains and AI systems themselves, where specific governments may leverage control over AI technologies exported by their national champions to gain strategic advantages or exert ingerence and influence over other nations. Balancing sovereignty with global cooperation and ensuring that AI technologies contribute to international peace and security remains a critical challenge for policymakers and stakeholders.

---

[1] For a brief introduction to the concept of "good digital sovereignty", see Belli, L. (June 2023). Building Good Digital Sovereignty through Digital Public Infrastructures and Digital Commons in India and Brazil. G20's Think20 (T20). For a comprehensive analysis of the various facets of the concept, especially from a Global South perspective, see Jiang, Min and Belli, Luca (Eds). Digital Sovereignty from the BRICS Countries. How Brazil, Russia, India, China and South Africa Are Reshaping Digital Infrastructure, Data Governance, and Online Service Regulation. Cambridge University Press. (2024).

[2] *Idem.*

[3] See Jiang, Min and Belli, Luca (Eds). *Supra*.

Second, AI transparency is a concept fraught with ambiguity, requiring clarification both in terms of substance and formal requirements to be meaningful and allow for accountability. Questions arise regarding the scope of transparency, for whom and about what AI transparency should apply, and how it should be operationalised. This volume provides some important answers to these questions. Without clear definitions and standards able to translate normative provisions into technical instructions for developers, thus facilitating both compliance and oversight, achieving meaningful transparency in AI systems becomes challenging, if not impossible, jeopardising enormously all accountability and trust-building efforts.

Third, accountability in AI governance suffers from a clear oversight deficit, making it problematic to ensure that AI systems or, more appropriately, those who develop, implement, and deploy them, can be held responsible by individuals, society, and regulators for the problems AI systems may cause. While traditional accountability mechanisms exist, such as legal liability and regulatory enforcement, applying them to AI systems poses unique challenges due to their complexity and autonomy of the systems, matched by the widespread incapacity – shared by most individuals, societies, and regulators – to clearly understand their functioning and foresee their impacts. Moreover, the distributed nature of AI technologies, involving multiple stakeholders in their value chain further complicates efforts to establish clear lines of accountability. Overcoming governance gaps and developing concrete mechanisms for AI accountability to stakeholders remain formidable challenges in the pursuit of responsible AI governance.

Conspicuously, addressing these challenges requires collaborative efforts from governments, industry, academia, civil society, and international organizations to develop clear norms, standards, regulatory frameworks, and oversight mechanisms that balance sovereignty, transparency, and accountability considerations while promoting innovation and protecting fundamental rights. Only through sustained engagement and cooperation can we ensure that AI technologies serve the common good and contribute to a more equitable and sustainable future.

24.2. Fostering Multistakeholder Cooperation

As we conclude our exploration of AI sovereignty, transparency, and accountability, it should be increasingly evident for the reader that addressing the complex and diverse challenges and realising the vast potential of AI technologies requires concerted efforts and collaboration across an ample spectrum of heterogeneous stakeholders. Multistakeholder cooperation, both in policy development and implementation, emerges as a fundamental imperative for achieving responsible and inclusive AI governance that aligns with societal values and aspirations. This is indeed one of the main reasons why the UN IGF Data and AI Governance Coalition, which promoted the elaboration of this volume, was established.

In the realm of policy development, the involvement of diverse stakeholders is essential to ensure that regulatory frameworks are comprehensive, balanced, and reflective of the diverse perspectives, interests and needs of all stakeholders. Governments, developers, enterprises, academics, and civil society representatives each bring unique expertise and insights to the table. Such multistakeholder participation can enrich and increase the quality of the research and policymaking processes, providing greater legitimacy and acceptance of AI regulations based on ample stakeholder consultation, and increasing the likelihood of creating effective

implementation mechanisms[4]. This volume showcases the potential brought about by such multistakeholder approaches, presenting a wide array of practical and theoretical frameworks for AI, from high-level regulation and control to firm-level governance.

Governments play a central role as regards the future of AI sovereignty, transparency, and accountability, being the ultimate setters of the policy agendas and creating an enabling environment for AI innovation while safeguarding public interests and values. However, governments alone are ill-prepared to effectively understand, develop and regulate AI. Effective policymaking requires close collaboration and consultation with non-governmental stakeholders from industry, academia, and civil society, which ultimately develop, study and are affected by AI systems, so that their expertise can be leveraged. Such cooperation is vital to fully understand existing risks, anticipate potential impacts, and address emerging challenges, and promote the development of AI that embeds the desired values and norms.

Developers and enterprises bring invaluable technical expertise and practical insights regarding the capabilities and limitations of AI technologies, what kind of incentive can promote its development and what type of restrictions are likely to stifle it, and what types of systems can allow embedding normative provisions into the design of AI. Their active participation in the policymaking process ensures that regulations are technically feasible, aligned with industry best practices, and conducive to innovation and competitiveness. However, one must not be naïve and remember that the natural role of private sector is not to promote human rights or sustainability but to pursue return on investment, maximising benefits and reducing costs.

Academics contribute cutting-edge research, ethical insights, and critical analysis to inform evidence-based policymaking and ensure that regulatory frameworks are grounded in rigorous scientific principles and legal and ethical considerations. In a similar way, civil society representatives, including advocacy groups, consumer organisations, and human rights defenders, serve as watchdogs and voices for marginalised communities, which are the most likely to be adversely affected by biased AI systems. Both academia and civil society must be supported and fully included in AI governance to ensure that the resulting AI regulations uphold fundamental rights, promote equity, and mitigate potential harms.

In the realm of policy implementation and oversight, effective multistakeholder cooperation is essential. Particularly, highly technical AI systems entail the need for regulators to cooperate with specialists and standardisation bodies to ensure that regulatory requirements are translated into technical standards that can be implemented effectively by industry stakeholders and understood by society at large.

Regulators, including government agencies and independent regulatory bodies, bear the duty to protect their citizens' fundamental rights, ensure healthy competition and defend national interests and sovereignty. Hence, to make sure that they can truly comply with their responsibilities, enforcing AI regulations, monitoring compliance, and addressing violations,

[4] For a deeper analysis of this perspective see e.g. Belli, L. (2015). A heterostakeholder cooperation for sustainable internet policymaking. Internet Policy Review, 4(2). https://doi.org/10.14763/2015.2.364 ; Belli, L. (2014). De la gouvernance à la regulation de l'Internet. Berger Levrault, Paris. (2015).

they need to cooperate with technical experts and be sufficiently well-resourced to develop appropriate implementation, and complement public regulation with technical standards.

In this perspective, technical standardisation bodies play a crucial role in developing technical standards and specifications that are vital for compliance with regulation, translating regulatory requirements into actionable guidelines for developers and enterprises. Close collaboration between regulators and standardisation bodies ensures that technical standards are aligned with regulatory objectives, reflect industry best practices, and facilitate (legal) interoperability[5] and compatibility among AI systems.

Moreover, multistakeholder engagement in the implementation and oversight process helps build trust, foster transparency, and enhance accountability in AI governance. By involving diverse stakeholders in monitoring compliance, conducting audits, and evaluating the effectiveness of regulatory measures, regulators and standardization bodies can ensure that AI technologies are deployed responsibly and ethically, in line with societal values and regulatory requirements.

In conclusion, fostering multistakeholder cooperation is essential for advancing responsible and inclusive AI governance that promotes the common good and addresses the complex challenges of the digital age. By working together, governments, developers, enterprises, academics, civil society representatives, regulators, and standardization bodies can harness the transformative potential of AI technologies while safeguarding human rights, promoting equity, and ensuring accountability and transparency in AI governance.

24.3. Challenges for Multistakeholder Cooperation in AI Governance

As emphasised above, multistakeholder cooperation plays an instrumental role enabling effective AI governance, particularly in addressing the complex challenges surrounding AI sovereignty, transparency, and accountability. However, achieving meaningful cooperation among diverse stakeholders presents significant challenges, stemming from highly divergent interests, the absence of a global framework for cooperation, and the difficulty of creating effective mechanisms for collaboration.

First, stakeholders in AI governance, including governments, developers, enterprises, academics, and civil society representatives, often have divergent interests and priorities. For

[5] Interoperability is usually described as “the ability to transfer and render useful data and other information across systems, applications, or components”. See International Telecommunication Union, ' GSR discussion paper: Interoperability in the digital ecosystem' (ITU 2015). Interoperability is therefore the property enabling the exchange and use of information across heterogeneous technologies and systems. Like technical interoperability, legal interoperability stimulates the exchange of information within different systems. As such, shared rules and principles amongst various juridical systems have the potential to reduce transaction costs, deflating barriers to cross-border trade, and foster non-measurable benefits, such as the protection of fundamental rights. See Weber, R. 'Legal Interoperability as a Tool for Combatting Fragmentation' (2014). (4) Global Commission on Internet Governance Paper Series; Belli L. and Doneda D. Data Protection in the BRICS Countries: Legal Interoperability through Innovative Practices and Convergence. (2023) 13 International Data Privacy Law; Belli, L. and Zingales, N. (2023). "Interoperability to foster open digital ecosystems in the BRICS countries". in Chinese Academy of Cyberspace Studies, Xinhua Institute, China Institute of International Studies. Shared Vision for the Digital World: Insights from Global Think Tanks on Jointly Building a Community with a Shared Future in Cyberspace. The Commercial Press.

instance, while developers and enterprises may prioritise market incentives and profit motives, civil society representatives may advocate for human rights, equity, sustainability, and social justice. Bridging these divergent interests and finding common ground can be challenging, particularly when market incentives conflict with human rights incentives, leading to tensions and disagreements in the policymaking process.

Second, the absence of a comprehensive global framework for cooperation in AI governance exacerbates challenges related to multistakeholder collaboration. Unlike other domains, such as trade or climate change, there is no universally accepted set of norms, principles, or institutions governing AI governance at the global level. As a result, coordination efforts are fragmented, inconsistent, and often driven by regional or national interests, hindering the development of cohesive and harmonized approaches to AI sovereignty, transparency, and accountability.

Consequently, the establishment of effective mechanisms for multistakeholder cooperation in AI governance is inherently complex and hard to achieve. It requires not only designing inclusive, transparent, and participatory processes that enable meaningful engagement and decision-making by all stakeholders, but also being able to persuade said stakeholders to work together in good faith. Achieving consensus among diverse stakeholders with competing interests and perspectives is often difficult, leading to delays, deadlock, and suboptimal outcomes in the policymaking process. Lastly, ensuring that mechanisms for cooperation are equitable, inclusive, and representative of all stakeholders' voices poses additional challenges, particularly in contexts characterised by considerable power imbalances and unequal access to resources and influence.

To overcome these challenges, concerted efforts and innovative approaches are needed. A commendable initiative in this sense is the recently established UN Advisory Body on AI, which aims at fostering dialogue, building trust, and promoting collaboration among diverse stakeholders. However, the practical success of this body is far from guaranteed. In such complex scenario, creating platforms and forums for ongoing dialogue and engagement among stakeholders, providing opportunities for knowledge sharing, consensus-building, and conflict resolution is essential. The UN IGF of which the Data and AI Governance Coalition is part could be a powerful ally in achieving this task.

Another key objective must be the development of frameworks and tools for assessing and mitigating risks associated with divergent interests and conflicting incentives, enabling stakeholders to identify common ground and align their objectives. This can, in turn allow to strengthen national, regional, and international cooperation and coordination mechanisms, based on partnerships and alliances, or even treaties, to promote convergence and harmonisation of AI governance approaches across different jurisdictions.

Last but not least, empowering marginalised and underrepresented stakeholders, including civil society organizations, indigenous communities, and vulnerable populations, to participate meaningfully in AI governance processes is vital for such processes to be truly legitimate and ensure that their perspectives and interests are adequately represented and addressed. Overcoming these challenges requires commitment, creativity, and collaboration among all stakeholders, with a shared vision of harnessing the transformative potential of AI technologies. It clearly also requires financial and intellectual resources to be able to uphold fundamental rights, promoting equity, and enhancing trust in AI governance, while also fostering its sustainable development.

24.4. Conclusion: Leveraging Standards for Translating AI Norms into Technical Specifications

In the rapidly evolving landscape of AI governance, one of the key challenges lies in translating abstract principles and regulatory frameworks into concrete technical specifications that developers can understand and implement effectively. Technical standards play a crucial role in bridging this gap by providing a common language and framework for translating AI norms and regulations into actionable guidelines and requirements that can be incorporated into AI systems and technologies.

The development of technical standards for AI sovereignty, transparency, and accountability is essential for promoting responsible AI development and deployment while ensuring compliance with regulatory requirements and human rights principles. These standards serve as a blueprint for designing, implementing, and assessing AI systems in accordance with established norms, thereby enhancing transparency, accountability, and trustworthiness in AI technologies.

By providing clear and actionable guidance on key aspects of AI governance, such as data protection, algorithmic transparency, and human oversight, technical standards enable developers to design AI systems that align with societal values, legal requirements, and ethical principles. They facilitate the integration of best practices and risk mitigation strategies into the development lifecycle of AI systems, helping to address concerns related to bias, fairness, and accountability.

Moreover, technical standards promote interoperability and compatibility among AI systems, enabling seamless integration and collaboration across different platforms and domains. They facilitate knowledge sharing, collaboration, and innovation in the AI community, driving continuous improvement and advancement in responsible AI development practices.

Overall, the development and adoption of technical standards for AI sovereignty, transparency, and accountability are essential for promoting ethical AI governance, fostering innovation, and building public trust in AI technologies. By providing a common framework for translating AI norms into technical specifications, these standards empower developers to build AI systems that are transparent, accountable, and aligned with societal values and regulatory requirements.

This volume has contributed in this sense, by providing frameworks grounded on strong practical and theoretical bases. Starting from the KASE framework for AI sovereignty, which paints a wider public policy and strategy picture, the authors have approached the issue of AI sovereignty, transparency and accountability from diverse perspectives that provide concrete steps toward AI implementation. These frameworks, of either narrower or wider application, are an invaluable source of inspiration for multistakeholder debates on what could constitute a general **Multistakeholder Framework for AI Sovereignty, Transparency, and Accountability.** The Data and AI Governance Coalition and UN IGF stakeholders could further explore this as a testament to the collective efforts and aspirations to assert good AI sovereignty, harnessing the transformative power of AI for the benefit of humanity while upholding the values of transparency, accountability, and justice for all.

References

BELLI, L. A heterostakeholder cooperation for sustainable internet policymaking. Internet Policy Review, v. 4, n. 2, 2015. Available at: https://doi.org/10.14763/2015.2.364

BELLI, L. Building Good Digital Sovereignty through Digital Public Infrastructures and Digital Commons in India and Brazil. G20's Think20 (T20), junho de 2023.

BELLI, L. De la gouvernance à la regulation de l'Internet. Paris: Berger Levrault, 2014.

BELLI, L.; DONEDA, D. Data Protection in the BRICS Countries: Legal Interoperability through Innovative Practices and Convergence. International Data Privacy Law, v. 13, 2023.

BELLI, L.; ZINGALES, N. Interoperability to foster open digital ecosystems in the BRICS countries. In: Chinese Academy of Cyberspace Studies, Xinhua Institute, China Institute of International Studies. Shared Vision for the Digital World: Insights from Global Think Tanks on Jointly Building a Community with a Shared Future in Cyberspace. The Commercial Press, 2023.

JIANG, M.; BELLI, L. (Eds). Digital Sovereignty from the BRICS Countries. How Brazil, Russia, India, China and South Africa Are Reshaping Digital Infrastructure, Data Governance, and Online Service Regulation. Cambridge University Press, 2024.

WEBER, R. Legal Interoperability as a Tool for Combatting Fragmentation. Global Commission on Internet Governance Paper Series, v. 4, 2014.