PART 1: LOCAL APPROACHES TO GLOBAL PROBLEMS

1. AI FROM THE GLOBAL MAJORITY: WHAT ARE WE DEBATING AND WHY?

LUCA BELLI AND WALTER BRITTO GASPAR

Abstract. The first Annual Report of the UN IGF Data and Artificial Intelligence Governance (DAIG) Coalition, released at IGF 2023, focused on "The Quest for AI Sovereignty, Transparency, and Accountability." Building on this outcome, the DAIG Coalition initiated a multistakeholder effort to discuss "AI from the Global Majority," aiming to provide insights for IGF 2024. This volume, compiled from an open Call for Essays, highlights AI initiatives from the perspectives of populations in Africa, Asia, Latin America, and the Middle East. These regions, often underrepresented in AI governance discussions, share a history of colonial exploitation and face ongoing neo-colonial and digital colonialism dynamics, which are becoming particularly evident as regards their adoption and regulation of AI as well as their capacity to contribute to global AI fora. The essays emphasise the need for inclusive and equitable representation in global AI dialogues. They explore the impact of AI on civil, political, economic, and social rights, addressing issues like surveillance, labour displacement, and environmental degradation. The volume advocates for AI systems designed with diverse data sets and inclusive practices to mitigate biases and promote fairness. Importantly, it outlines not only problems faced by the global majority, but also relevant solutions emerging from such countries.

INTRODUCTION

This volume presents the results of the 2024 works undertaken by the Data and Artificial Intelligence Governance (DAIG) Coalition[1] established under the auspices of the United Nations Internet Governance forum (IGF). The Coalition is a multistakeholder group aimed at fostering discussion of existing approaches to data and AI governance, promoting analysis of good and bad practices to identify what solutions should be replicated and which ones should be avoided by stakeholders to achieve a sustainable an effective data and AI governance.

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

After having successfully released at the IGF 2023 the first Annual Report of the Coalition, featuring analyses from 34 authors dedicated to "The Quest for AI Sovereignty, Transparency and Accountability"[2], the DAIG Coalition has promoted a multistakeholder effort aimed at discussing "AI from the Global Majority", to provide valuable inputs that could feed into IGF 2024 discussions and beyond. Authors of this volume responded to an open Call for Essays, shared over the DAIG mailing

[1] For further information about the DAIG Coalition of the UN Internet Governance Forum, see https://intgovforum.org/en/content/dynamic-coalition-on-data-and-artificial-intelligence-governance-dc-daig

[2] Belli, L. and Gaspar, W.B. (Eds.) The Quest for AI Sovereignty, Transparency and Accountability. Official Outcome of the UN IGF Data and Artificial Intelligence Governance Coalition. FGV. (2023). https://hdl.handle.net/10438/34295

list, with the aim to collect valuable insight analysing AI initiatives from the perspectives of global majority populations. Indeed, most DAIG members felt that such perspectives are often underrepresented in discussions of data and AI governance, as noted in a dedicated multistakeholder workshop organised by DAIG Coalition members during the Computers Privacy and Data Protection Conference Latin America (CPDP LatAm) 2024.[3]

The term "global majority" refers to the populations of an ample range of highly heterogenous countries from Africa, Asia, Latin America, and the Middle East, which together make up most of the world's population. This concept challenges the traditional Eurocentric perspective, highlighting the importance of recognising and valuing the diverse range of cultures, histories, and contributions from the abovementioned regions.

Importantly, despite their heterogeneity, almost all countries and populations from the global majority share the feature of being former colonies of global north countries, which implemented extractive practices, concentrating resources and violently subjugating local populations for multiple centuries, thus considerably contribution to the enormous inequalities that still characterise these countries. Importantly, this volume starts form the assumption that some of these exploitative practices merely evolved into neo-colonial dynamics, while other new forms of digital colonialism have emerged over the past decades.[4] As a result, global south countries find themselves in a very thorny situation, trying to shape their AI approaches while being in a situation of clear dependency and lack of essential transparency and accountability tools, which are necessary to elaborate solid strategies, policies and regulations.[5]

By focusing on the global majority, we emphasise the need for understanding the point of views, the needs and the perspective of this global majority of countries, in a perspective of inclusivity and equitable representation in global dialogues and decision-making processes. This shift in perspective is crucial for addressing global challenges in a way that is fair and just for all.

1.1. AI AND HUMAN RIGHTS IN THE CONTEXT OF THE GLOBAL MAJORITY

Artificial Intelligence (AI) significantly impacts civil and political rights, especially for the global majority. Surveillance technologies powered by AI can infringe on multiple fundamental rights, especially privacy, data protection, and freedom of expression. In many countries, these technologies are used to monitor and suppress dissent, disproportionately affecting marginalised communities. For instance, facial recognition systems, often less accurate for people of color, can lead to wrongful arrests and increased surveillance of minority groups. Additionally, AI in law enforcement and judicial systems can perpetuate existing biases, resulting in unfair treatment and discrimination. In this context, it is essential to develop and implement AI technologies that respect and protect civil and political rights, ensuring that they do not become tools of oppression.

---

[3] See the report of the CPDP LatAm 2024 session on "AI from the Global Majority: Meeting of the UN IGF Coalition on Data and AI Governance" available at https://cpdp.lat/en/

[4] Anibal Quijano. (2000) 'Coloniality of Power, Eurocentrism, and Latin America', Nepantla: Views from South, 1(3), pp. 533– 580; Everisto Benyera. (2021). The Fourth Industrial Revolution and the Recolonisation of Africa: The Coloniality of Data. Routledge.

[5] Such consideration is corroborated by the general lack of AI sovereignty in most countries, as highlighted in the 2023 Outcome of the DAIG Coalition. See *supra* n. (2).

Furthermore, AI's influence extends to economic and social rights, where it can both create opportunities and exacerbate inequalities. Automation and AI-driven technologies have the potential to transform industries and create new jobs. However, they also pose a risk of labour displacement, particularly in regions where economies are heavily reliant on low-skilled labour. For example, the introduction of AI in manufacturing and agriculture can lead to job losses for workers who lack the skills to transition to new roles.

Access to AI technologies and the benefits they bring is often uneven, further widening the gap between the global majority and more developed nations. To address these challenges, it is crucial to invest in education and training programs that equip workers with the skills needed for the AI-driven economy.

As an instance, the rise of AI and automation has significant implications for labour markets, particularly in the global majority. While AI can enhance productivity and create new job opportunities, it can also lead to labour exploitation. Workers in low-wage jobs may face increased pressure and job insecurity as companies adopt AI technologies to cut costs. For instance, gig economy workers may be subjected to algorithmic management practices that prioritize efficiency over worker well-being. It is important to develop policies and regulations that protect workers' rights and ensure fair labour practices in the AI-driven economy.

As this volume illustrates, AI systems can perpetuate exclusion and discrimination if not designed and implemented with inclusivity in mind. Biases in data and algorithms can lead to discriminatory outcomes, affecting access to services, employment, and justice. For instance, biased hiring algorithms can disadvantage candidates from certain backgrounds, while biased credit scoring systems can limit access to financial services for marginalised communities. The lack of representation of the global majority in AI development exacerbates these issues, as the perspectives and needs of these populations are often overlooked. It is essential to ensure that AI systems are developed with diverse data sets and inclusive practices to mitigate these risks.

The underrepresentation of the global majority populations, interests, and perspectives in AI research and development means that their unique challenges and contexts are not adequately addressed. This lack of diversity in the tech industry leads to the creation of AI systems that do not serve the needs of all users equally, reinforcing existing inequalities. For example, language models trained primarily on English data may not perform well for speakers of other languages, limiting their accessibility and usefulness. Increasing the representation of the global majority in AI development is crucial for creating technologies that are equitable and inclusive.

Furthermore, global majority populations are likely to be the ones suffering g the most from the environmental impact of AI infrastructure, which is increasingly recognised as a critical issue. The energy consumption of AI systems, particularly those requiring large-scale data processing and storage, contributes to carbon emissions and environmental degradation. This impact is felt disproportionately in regions already vulnerable to climate change, many of which are part of the global majority. For example, data centres in developing countries may strain local energy resources and contribute to pollution. To mitigate these effects, it is essential to develop energy-efficient AI technologies and promote sustainable practices in the tech industry.

The following section provides an overview of how the issues are explored in this volume, illustrating the complexities that the global majority is facing but also the solutions that are emerging from these countries.

1.2. How is this volume addressing these issues?

This volume examines the transformative impact of Artificial Intelligence (AI) on contemporary societies. Papers are organised around five thematic axes that provide a structured exploration of those impacts and proposed frameworks and solutions to existing issues, with contributions from diverse regional and global perspectives focused on some of the key challenges which are particularly relevant for the Global South.

The first section, **Local Approaches to Global Problems**, delves into how nations tailor AI-driven solutions to address unique domestic challenges while engaging with broader global trends. Luca Belli's paper, "AI Meets Cybersecurity: A Brazilian Perspective", evaluates the role of AI in cybersecurity both as defensive and an offensive tool. Belli situates the discussion in Brazil and advocates for integrated multistakeholder governance frameworks through the creation of a "Brazilian Cybersecurity and Digital Transformation System". Subsequently, Sizwe Snail and colleagues offer a constructive critique of South Africa's AI governance landscape in "The Law on Artificial Intelligence in South Africa". Their analysis of the South African Draft AI Strategy (SADAIS) and the National AI Policy Framework highlights the urgency of aligning national policies with evolving regional and global legal standards – an effort that, as the authors demonstrate, is still lacking in the country.

Additionally, Zijing Liu et al. provide an empirical study of China's judicial innovations in "Building Smart Courts Through Large Legal Language Models". Their work reflects on the current landscape of AI-enabled legal decision-making in Chinese courts, drawing lessons from regional variations in smart court implementation. Finally, Nils Brinker and Richard Skalt, in "Fox Guarding the Chickens", expose inherent biases in the EU AI Act's risk management obligations, emphasizing the critical need for impartial mechanisms to address third-party risks. This points to relevant blind spots in the European model of AI regulation – an important point to be considered by Global South countries amidst the so-called "Brussels effect" on various regulatory fronts, indicating that, while learning from the achievements and shortcomings of the European experience is useful, an original and context-adequate regulation is crucial for these countries.

In the second section, **The Emergence of Regional Solutions**, the focus shifts to regional strategies that navigate AI's complexities within distinct socio-political contexts. Pablo Trigo Kramcsák et al. analyse the "Incipient Latin American Approach to AI Governance", showcasing how Latin American countries have been establishing their own AI frameworks influenced by EU regulatory principles. These are early-stage efforts and face significant challenges, especially in promoting regulatory coordination between existing and new authorities. Chinasa T. Okolo's "RICE Governance Framework" proposes a cohesive strategy for African nations, emphasizing the need to reform governance measures, integrate regulatory efforts, improve regional cooperation, and boost enforcement. Andrea Bauling explores the legal challenges of AI in education in "AIED and Student Data Privacy in Africa", highlighting the need to craft Africa-centric policies focused on AIED that protect student data while fostering technological innovation. Ekaterina Martynova contributes a commentary on the Council of Europe Framework Convention on AI and Human Rights, which provides, through its soft regulatory approach, common ground and a first step toward international regulatory approaches – a model that

could inspire the BRICS. Finally, Yonah Welker introduces a human-capacity-centred AI policy emphasizing disability inclusion and emphasizes the need for disability representation in policy geared toward AI.

The third section, **Global Majority Facing AI**, centres on equity and justice for the Global Majority, addressing exploitation and systemic inequalities perpetuated by AI. Elise Racine's "Reparative Algorithmic Impact Assessments" outlines a justice-oriented framework for mitigating the harms of AI-powered systems through an approach that combines culturally sensitive participatory methods and a reparative praxis and decolonial, Intersectional principles. Alice Rangel Teixeira challenges the ethical foundations of mainstream AI principles in "AI Ethics for the Global Majority", proposing decolonial feminist bioethics as an alternative approach focused on power relations, relational autonomy, shared responsibility, empirical evidence, and local contexts.

This theme continues with analyses of content moderation harms and how they disproportionately affect Global Majority communities, with Zeerak Talat and Hellina Hailu's "Exploitation all the way down: Calling out the root cause of bad online experiences for users of the 'majority world'"; and the socio-political implications of false information, compounded by generative AI technologies, as well as the way forward in Isha Shuri and Shiva Kanwar's "Countering False Information: Policy Responses for the Global Majority in the Age of AI".

Richard Ngamita's contribution, "Addressing the Challenges of AI Content Detection in the Global South", explores the limitations of existing AI systems in detecting manipulated media, particularly cheap fakes, and advocates for the development of AI models trained on local data, alongside inclusive content moderation policies, to safeguard civic participation and democracy in the Global South. This axis is closed by Guangyu Qiao-Franco and Mahmoud Javadi's "Bridging the gap between the North and South in the governance of dual-use artificial intelligence technologies", on the implications of dual-use AI technologies, highlighting the critical need for equitable global AI governance.

In **Social Challenges of AI**, the fourth axis, the discussion broadens to include labour, education, and environmental sustainability. A case study from MIT Critical Data by Catherine Bielick, Rodrigo Gameiro and Leo Celi underscores the necessity of inclusive AI development practices and how to achieve them, while Avantika Tewari critiques various aspects of the relations between platforms and users and how conceptualizing data subjects as prosumers reinforces issues related to participation and labour and to the effective control over personal data.

Papers on healthcare and education explore how AI can address, but also exacerbate, existing inequities. Amrita Sengupta and Shweta Mohandas, in "Cost or Benefit? The Impact of AI on the Work of Medical Practitioners", analyse the integration of AI into healthcare practices, focusing on its current use and impact on medical workflows in India. Through primary research, the authors highlight both the potential benefits and the challenges for medical professionals. Faizo Elmi's "Reimagining Education: Potential Solutions for Nomads", explores how AI technologies, such as adaptive learning platforms and virtual classrooms, can address the educational challenges faced by nomadic populations, highlighting the need to overcome barriers like technological infrastructure and cultural adaptation to ensure equitable and effective implementation.

Finally, Jess Reia, Rachel Leach, and Anuti Shah, in "The Need for Transnational Perspectives on the Social, Legal and Environmental Impact of Artificial Intelligence", argue for integrating environmental

justice into AI regulatory frameworks. By examining cases in the US and Brazil, they highlight the geopolitical and ecological costs of AI development and propose incorporating hidden costs especially affecting marginalised and global majority communities into the transnational regulatory ecosystem.

The final section, **Foresighted Solutions for Present Problems**, offers innovative approaches to pressing AI-related challenges. Matheus Alles examines the ontological shifts at the intersection of law and data science, advocating for a reflexive legal rationality. This would allow for effective and ethical integration of data science in the legal field, through a process of critical assessment of the legal community and adaptation to new forms of knowledge and rationality.

Julio Gabriel Mercado, in "People-Centered Justice AI: Data Dimensions for Embracing a Responsible Digital Transformation", discusses how effective digital transformation of justice must go beyond mere technological adoption, incorporating Open Justice principles such as transparency, accountability, and public participation. Liisa Janssens, Saskia Lensink, and Laura Middeldorp, in "Fostering AI Research and Development: Towards a Trustworthy LLM", discuss compliance challenges and ethical considerations in the development of Large Language Models (LLMs) through a scenario-based analysis, focusing on the implications of including or omitting an opt-out option for personal data removal. Finally, Ronald Musizvingoza examines the potential of using synthetic data to create representative datasets that reflect diverse gender experiences while addressing the risks of bias and misuse. Collectively, these contributions emphasize the importance of aligning technological advancements with ethical imperatives and human-centred design principles.

By weaving together these thematic axes, this volume provides a comprehensive understanding of AI's transformative role in society. It not only highlights the potential benefits of AI but also critically engages with its risks, especially in the face of Global Majority communities and countries, advocating for a context-specific and balanced perspective for inclusive, equitable, and sustainable AI governance.

2. AI MEETS CYBERSECURITY: A BRAZILIAN PERSPECTIVE ON INFORMATION SECURITY AND AI CHALLENGES

LUCA BELLI

Abstract. Artificial Intelligence (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. This paper starts by exploring the distinction between defensive AI and offensive AI in the context of cybersecurity. Subsequently, it focusses on the Brazilian context to explore how the country is dealing with the emerging threats and opportunities presented by the intersection of AI and Cybersecurity. Lastly, it puts forward some concise recommendations for policymakers advocating for multistakeholder cooperation to be embedded in the future Brazilian Cybersecurity Strategy and Brazilian Cybersecurity Agency, to cope with the increasing complex intersection between cybersecurity and AI. Ideally, such recommendations could be integrated in the proposals for a new strategy and agency that will be issued by the new National Cybersecurity Committee (known as "CNCiber") a multistakeholder advisory body recently established by the Brazilian Presidency.[6]

INTRODUCTION

Artificial Intelligence (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[7], autonomous vehicle-management, etc. – creates numerous new, and sometimes unpredictable, risks and can open new avenues in attack methods and techniques.[8] 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.

---

[6] An early version of this paper was presented at the Digital Democracy Network (DDN) Conference organized by the Carnegie Endowment in June 2024, to be published in the Carnegie Endowment collection on digital technology and democracy. The author would like to sincerely thank Steven Feldstein, organizer of the conference and editor of the collection, as well as the DDN participants for their comments and feedback on the initial version of this paper.

[7] L. F. Salomão. Artificial Intelligence: technology applied to conflict management within the Brazilian judiciary. FGV. (2022). https://hdl.handle.net/10438/33954 ; L. Belli et al. Courting AI: How Brazilian Courts are Using and Regulating AI. in A. Limante and M. Zalnieriute. Cambridge Handbook of Courts and AI. (forthcoming).

[8] Belli et al. Cibersegurança: uma visão sistêmica rumo a uma proposta de Marco Regulatório para um Brasil digitalmente soberano. FGV. (2023). https://hdl.handle.net/10438/33784 ; ENISA. AI Cybersecurity Challenges Threat Landscape for Artificial Intelligence. (2020). https://www.enisa.europa.eu/publications/artificial-intelligence-cybersecurity-challenges

This paper adopts the definition of AI system offered by article 4 of the latest version of Bill 2338/2023, which is largely based on the definitions offered by the EU AI Act and the OECD[9], and therefore highly unlikely to be altered. The definition proposed by the Brazilian bill reads as follows:

> *"artificial intelligence (AI) system: a machine-based system that, with different degrees of autonomy and for explicit or implicit purposes, infers, from a set of data or information it receives, how to generate results, in particular, prediction, content, recommendation or decision that can influence the virtual, physical or real environment."*[10]

Importantly, the dual nature of AI allows to utilise such technology to both strengthen and undermine cybersecurity. However, while both the aforementioned Brazilian AI Bill and other leading examples of AI regulation such as the EU AI Act consider cybersecurity a key concern for the development and deployment of AI systems, neither offers clear guidance on how to concretely assess risks and implement regulation cybersecurity aspects of IA.

In this respect this paper argues that considerable work is needed to support the implementation of existing and proposed frameworks, particularly through the adoption of technical standards able to specify and give meaning to highly vague formulations, that are typically adopted by AI regulatory frameworks to define cybersecurity risk management provisions.

First, this paper explores the distinction between defensive AI and offensive AI in the context of cybersecurity. Second, it focusses on the Brazilian context to explore how the country is dealing with the emerging threats and opportunities presented by the intersection of AI and Cybersecurity and what type of provisions are dedicated to the issue in the proposed AI Bill. Lastly, it puts forward some concise recommendations for policymakers.

2.1. AI AND CYBERSECURITY: A COMPLICATED RELATIONSHIP

The relationship between AI and cybersecurity is based on how the former is used to impact the latter and vice versa, and the resulting defensive, offensive, or adversarial capabilities.[11] While there is already a conspicuous body of research on the technical aspects of AI and cybersecurity, it is surprising that remarkably scarce research exists on the interactions of AI and cybersecurity from a regulatory and governance angle. This essay aims at understanding what is at stake when we adopt this latter angle and policy issues should be considered as priorities.

To do so, we should initially distinguish between defensive AI and offensive AI. Defensive AI usually leverages machine learning and other AI techniques to enhance the cybersecurity and resilience of computer systems, networks, and data bases, and to protect individuals, shielding them against cyber threats.[12] In this perspective, AI systems can increase the effectiveness of security controls aimed at

---

[9] S. Russell, K. Perset, M. Grobelnik. Updates to the OECD's definition of an AI system explained.OECD.AI Policy Observatory. (29 November 2023). https://oecd.ai/en/wonk/ai-system-definition-update

[10] Bill No. 2,338, of 2023 on the development, promotion, ethical and responsible use of artificial intelligence based on the centrality of the human person. https://www25.senado.leg.br/web/atividade/materias/-/materia/157233

[11] L. Belli *et al*. (2023). *Cit. supra*. N (8).

[12] B. Geluvaraj; P. M Satwik; and T.A. Ashok Kumar. The future of cybersecurity: Major role of artificial intelligence, machine learning, and deep learnin in cyberspace. International Conference on Computer Networks and Communication Technologies: ICCNCT 2018. Springer. (2019).

protecting specific assets, for instance through automated malware analysis, active firewalls, automated cyber threat intelligence operations, etc.[13]

In contrast, offensive AI, also known as AI-powered cyberattacks, involves the use of AI to launch malicious activities, enhancing vulnerability detection and exploitation, developing new cyberattacks types and strategies or automating the exploitation of existing vulnerabilities. Lastly, we should mention the that adversarial AI is a subcategory of offensive AI and refers to the manipulation of AI systems to cause them to make incorrect predictions. This can be achieved by tampering with the input data or poisoning the training data used to develop the AI system. [14]

Importantly, this papers adopts the definition of cybersecurity provided by the Telecommunication Standardization Sector of the International Telecommunication Union (ITU-T), which is noteworthy for being a rare example of consensual definition at the international level, according to which:

> *"Cybersecurity is the collection of tools, policies, security concepts, security safeguards, guidelines, risk management approaches, actions, training, best practices, assurance and technologies that can be used to protect the cyber environment and organization and user's assets. Organization and user's assets include connected computing devices, personnel, infrastructure, applications, services, telecommunications systems, and the totality of transmitted and/or stored information in the cyber environment. Cybersecurity strives to ensure the attainment and maintenance of the security properties of the organization and user's assets against relevant security risks in the cyber environment[15]."*

The amplitude provided by the above definition tellingly illustrates the complexity of cybersecurity and the necessity of adopting a collaborative and coordinated multistakeholder approach to the issue, as no single stakeholder can guarantee cybersecurity in a vacuum.

2.2. A PARADIGM SHIFT

The integration of AI capabilities has constituted a watershed moment in the development of cyber threats, significantly augmenting the efficacy, scope, scale, and precision of malicious cyber operations. This evolution marks a paradigm shift in the cybersecurity landscape, fundamentally altering in multiple ways the nature of both offensive and defensive strategies.

First, the democratisation and increased sophistication of AI tools enables cybercriminals to automate and refine their attacks, making them more effective, callable, dynamic, and difficult to detect. Machine learning algorithms, for instance, can analyse vast amounts of data to identify vulnerabilities in systems and networks, enabling attackers to exploit these weaknesses with greater precision. Automated phishing campaigns can be tailored to individual targets based on data harvested from social media and other sources.

This personalisation increases the likelihood of success, as the messages appear more convincing and relevant to the recipient. Critically, concern about AI-enhanced malicious attacks now represents the

---

[13] L. Belli *et al*. (2023). *Cit. supra*. N (8).

[14] M. Malatji; A. Tolah. Artificial intelligence (AI) cybersecurity dimensions: a comprehensive framework for understanding adversarial and offensive AI. AI and Ethics (2024).

[15] ITU-T. (2009). Recommendation X.1205 (04/08): Overview of cybersecurity. Approved in 2008-04-18. https://www.itu.int/rec/T-REC-X.1205-200804-I

top emerging risk according to the latest version of the periodic Gartner study dedicated to risk monitoring, due to "the relative ease of use and quality of AI-assisted tools, such as voice and image generation, increase the ability to carry out malicious attacks with wide-ranging consequences."[16]

Second, AI is likely to expand the scope of cyberthreats by allowing attackers to increase the scale of their operations with minimal human intervention. As an instance, AI-powered botnets can be used to operate massive Distributed Denial-of-Service (DDoS) attacks, able to overwhelm electronic networks. Ransomware attacks are also becoming more sophisticated and widespread due to AI support, leading to the consolidation of Ransomware-as-a-Service (RaaS) as a thriving industry with global range. In this context AI is sensibly lowering barriers to entry for attackers, increasing ease and availability of ransomware, via AI-driven malware capable of quickly and autonomously spread across networks, encrypt data, and demand ransoms, leading to high cost of recovery and downtime. [17]

Third, AI systems can substantially increase attackers' ability to analyse complex datasets and recognise patterns, thus allowing to execute highly targeted and precise attacks. For example, AI can be used to identify high-value targets within organisations and tailor attacks to their specific roles and responsibilities. AI can also allow cybercriminals to create realistic audio and video impersonations, that can be considered as "deepfakes", which can be used in social engineering attacks to manipulate individuals into divulging sensitive information or authorising fraudulent transactions.[18] It is now memorable the case of an elaborate deepfake scam, where a finance worker at a multinational firm was duped into paying USD 25 million to fraudsters who had lured him into a fake emergency call.[19]

Fourth, the increasing sophistication of deepfakes can be used to orchestrate disinformation campaigns for both financial and political purposes. These technologies pose a novel cybersecurity threat to of democratic processes by enabling malicious actors to undermine information integrity at an unprecedented scale. The current democratisation of AI implies much greater and easier access to AI systems that until just few years ago were only accessible to researchers and highly specialised companies or governmental actors. This process leads to an enormous expansion of the attack surface, both in terms of potential perpetrators and in terms of potential vulnerabilities and attack strategies that can be used.[20]

Importantly, AI-driven cyberattacks have acquired a dynamic nature, being able to adapt to changing defensive measures, making detection and mitigation more challenging. By using machine learning capabilities, attackers can alter malicious software in real time to avoid detection by traditional antivirus systems. For instance, AI-enhanced polymorphic or metamorphic malware is able to mutate

---

[16] Gartner. 2Q24 Emerging Risks Report. https://www.gartner.com/en/documents/5529395
[17] Hassan, S.M., Wasim, J.: Study of artificial intelligence in cyber security and the emerging threat of ai-driven cyberattacks and challenges. J. Aeronaut. Mater. 43(1), 1557–1570. (2023).
[18] MIT Technology Review Insights. Preparing for AI-enabled cyberattacks. (8 April 2021). https://www.technologyreview.com/2021/04/08/1021696/preparing-for-ai-enabled-cyberattacks/
[19] H. Cen and K. Magramo. Finance worker pays out $25 million after video call with deepfake chief financial officer.
CNN. (4 February 2024). https://edition.cnn.com/2024/02/04/asia/deepfake-cfo-scam-hong-kong-intl-hnk/index.html
[20] Lorenzo Pupillo et al. Artificial Intelligence and Cybersecurity Technology, Governance and Policy Challenges Final Report of a CEPS Task Force. Centre for European Policy Studies (CEPS) Brussels. (May 2021). https://www.ceps.eu/wp-content/uploads/2021/05/CEPS-TFR-Artificial-Intelligence-and-Cybersecurity.pdf

its features or automatically "recoding" itself when it propagates to evade pattern matching detection systems that are traditionally deployed as security solutions. Furthermore, AI systems can be used to quickly identify and exploit zero-day vulnerabilities before patches can be developed and deployed.[21]

Crucially, defenders are also increasingly employing AI-based systems to detect cyber threats and vulnerabilities and rapidly respond e.g. leveraging AI to identify software bugs and self-patch them. However, within a sort of cybersecurity arms race, attackers are also leveraging AI to outmanoeuvre these defences. This situation where both sides continuously refine their techniques, defensive AI systems must evolve rapidly to detect new attack patterns and anomalies, while policy and governance framework must be crafted to mitigate risks and facilitate communication, collaboration and coordination amongst cybersecurity stakeholders.

2.3. Understanding the Brazilian Context

Despite relevant advancements in recent years, the regulation of AI and cybersecurity in Brazil is highly fragmented, limited and poorly implemented. Due to the adoption of multiple sectoral regulations dedicated to cybersecurity, Brazil has climbed several rankings,[22] but the regulatory oversight and cybersecurity implementation remains patchy, being the responsibility of many different and uncoordinated entities, including sectoral regulators, private and public Computer Security Incident Response Teams, and the military.[23]

While Brazil does not have a general cybersecurity law, the top institution responsible for cybersecurity governance and policy proposal is the Institutional Security Cabinet (GSI in its Portuguese acronym) of the Brazilian Presidency. However, the GSI remit is limited to the federal administration thus limiting enormously the scope of its reach. Importantly, in December 2023, Brazil adopted a new National Cybersecurity Policy and established a new multistakeholder National Cybersecurity Committee[24] (known as "CNCiber"), of which the author of this paper has been appointed a member.[25] Amongst the tasks of CNCiber is the elaboration of a proposal for a new national cybersecurity strategy and a new body for cybersecurity governance and regulation.

Indeed, one of the reasons of the fragmented Brazilian approach to cybersecurity is the lack of a unique institution responsible for coordinating the various dimensions of it. At the same time, the previous Brazilian National Cybersecurity Strategy expired in December 2023. Hence, at the moment of this writing, Brazil does not have an actionable cybersecurity strategy allowing the country to organically

[21] ENISA (2020). *Cit. supra*. N (8).
[22] Most notably, in 2020 Brazil jumped up 53 positions, from 71st to 18th, in the Global Cybersecurity Index (GCI) elaborated by the International Telecommunications Union. In the Americas region, Brazil reached the 3rd position, surpassed only by the USA and Canada. The 2024 edition of the GCI considers Brazil as a "Tier 1 – Role-modelling" country. https://www.itu.int/epublications/publication/global-cybersecurity-index-2024
[23] L. Belli *et al*. (2023). *Cit. supra*. N (8).
[24] Brazilian Presidency. Presidential Decree 11.856/2023 https://www.in.gov.br/en/web/dou/-/decreto-n-11.856-de-26-de-dezembro-de-2023-533845289#wrapper
[25] Brazilian Presidency. GSI Ordonnance 6/2025/ https://www.in.gov.br/en/web/dou/-/portaria-n-6-de-9-de-fevereiro-de-2024-542752145 ; CyberBRICS. Professor Luca Belli appointed member of the new Brazilian Cybersecurity Committee. (16 February 2024). https://cyberbrics.info/professor-luca-belli-appointed-member-of-the-new-brazilian-cybersecurity-committee/

tackle the multiple – and mounting – cyberthreats it faces and assess the ways in which AI technologies are impacting such threats.

Furthermore, only limited AI regulation exist, falling primarily under the purview of the Brazilian Data Protection Authority, ANPD in its Portuguese acronym. In this context, the Brazilian National Congress is currently considering regulating AI with a dedicated framework, which would include cybersecurity obligations related to AI systems. While numerous AI bills are under consideration, Bill 2338/2023 seems to be the most complete and well-structured, being the result of multiple years of hearings and multistakeholder consultations. However, at the time of this writing, the Brazilian Congress has not adopted the Bill yet.

2.3.1. INFORMATION SECURITY?

Information security is an essential dimension, common to both AI and cybersecurity. In Brazil, the National Data Protection Authority (ANPD) is tasked with enforcing the Brazilian General Data Protection Law (LGPD) and ensuring that organisations comply with data protection obligations, including regarding the implementation of data security obligations. Data security is a fundamental principle set by the LGPD, aimed at ensuring that data is protected against unauthorised access, loss, alteration, damage, or destruction. Importantly, the LGPD explicitly establishes a security-by-design obligation for data controllers and processors, who need to implement security measures that the data subject "can expect", to demonstrate that personal data processing activities are regularly undertaken (Article 44).

According to Article 46 of the LGPD, "The processing agents must adopt security, technical, and administrative measures capable of protecting personal data from unauthorised access and from accidental or unlawful situations of destruction, loss, alteration, communication, or any form of inappropriate or unlawful processing."[26] In particular, indent 2 of this article highlights that information security measures "must be observed from the design phase of the product or service until its execution." Additionally, Article 49 specifies that "The systems used for processing personal data must be structured to meet security requirements, best practices, and governance standards, as well as the general principles provided for in this Law and other regulatory standards."[27]

To comply with the LGPD, processing agents are supposed to implement solid information security solutions. Such measures are suggested by ANPD in Orientation Guide[28] and include administrative measures, such as i) the definition of an information security policy; ii) awareness raising and capacity building; iii) and contract management; as well as the establishment of technical measures, such i) the establishment of access controls to ensure that only authorised individuals have access to personal data; ii) the use of security measures such as encryption to protect personal data during storage and transmission; iii) backup and recovery to ensure data availability in case of loss or damage; iv) and vulnerability monitoring and detection, to promptly identify and respond to data security breaches.

In practice, however, data security compliance is poor at best, given the total absence of ANPD oversight as regards this matter, in the first four years since its inception, despite the enormous and

[26] The Brazilian General Data Protection Law (LGPD) – Unofficial English Version. CyberBRICS. (2020). https://cyberbrics.info/brazilian-general-data-protection-law-lgpd-unofficial-english-version/
[27] Ibid.
[28] See ANPD. Orientation Guide on Information Security for Small-Scale Data Processing Agents. (October 2021). https://www.gov.br/anpd/pt-br/documentos-e-publicacoes/guia-vf.pdf

growing amount of information security incidents in Brazil. Indeed, the tropical giant ranks second globally for cyberattacks[29], which have exploded in number and sophistication due to the adoption of AI systems, making them more complex and difficult to detect, as exposed previously.

The ANPD is the body responsible for overseeing and regulating the implementation of LGPD, including regarding data security. However, so far, the ANPD has not regulated data security, despite having the possibility to do so, having simply adopted the above-mentioned Orientation Guide and a recent Regulation on the Communication of Information Security Incidents[30].

However, focusing on the communication of cybersecurity accidents rather than on overseeing the implementation of the existing information security obligation seems rather counterproductive. It is rather absurd to invest resources in overseeing the communication of the tragedy rather than on the implementation of the norms that would avoid or at least mitigate the tragedy itself. Furthermore, despite having a clear mandate to enforce the LGPD provisions on data security, no single sanction has been adopted so far for lack of compliance with such norms, in a country where the number of cybersecurity incidents is raising exponentially and only 2023 registered 103 Billion cyberattacks.[31]

A more interesting approach has been recently adopted by worth noting that the Ordinance SGD/MGI No. 852 of 28 March 2023 established the Privacy and Information Security Program (PPSI)[32] dedicated to enhance cybersecurity of the Brazilian public administration. Data governance in the Brazilian public sector is rather heterogeneous with most public administrations still having very basic cybersecurity governance despite the enormous digitalisation that Brazilian public services undertook since the Covid19 pandemic.[33] The PPSI program is therefore a welcome initiative, characterised by a set of projects and adaptation processes aimed at increasing cybersecurity maturity, resilience, effectiveness, collaboration, and intelligence.

The LGPD and PPSI should be considered as two essential information security pillars, but not sufficient on their own. In this context, it is essential that the future Cybersecurity Strategy, to be proposed by the National Cybersecurity Council, specify information security criteria for categories of sensitive information that are not personal.[34] Furthermore, it seems desirable that the future Brazilian Cybersecurity Agency establish cooperation agreements, and ideally a coordination mechanism, with

---

[29] Nakamura J. Brasil é vice-campeão em ataques cibernéticos, com 1.379 golpes por minuto, aponta estudo. CNN Brasil. (30 October 2024).

[30] ANPD. Resolution No. 15/2024. Approves the Security Incident Communication Regulation. (26 April 2024). https://www.in.gov.br/en/web/dou/-/resolucao-cd/anpd-n-15-de-24-de-abril-de-2024-556243024

[31] L. Belli *et al*. (2023). *Cit. supra*. N (8).

[32] The ordinance was issued by the Secretariat of Digital Government of the Ministry of Management and Innovation in Public Services. Further information on the program and can be found at https://www.gov.br/governodigital/pt-br/privacidade-e-seguranca/programa-de-privacidade-e-seguranca-da-informacao-ppsi

[33] L. Belli et al. Governança de dados no setor público: dados abertos, proteção de dados pessoais e segurança da informação para uma transformação digital sustentável. Lumen Juris. (May 2024). https://hdl.handle.net/10438/35341

[34] Such specification is utilised e.g. by the Chinese Cybersecurity Law and Data Security Law, which prescribe the adoption of specific measures to protect "important" or "core" data whose security is essential for the well-functioning of national critical infrastructure. See L. Belli L. "Cybersecurity Policymaking in the BRICS Countries: From Addressing National Priorities to Seeking International Cooperation" (2021) The African Journal of Information and Communication (AJIC), (28). doi:10.23962/10539/32208

the ANPD as well as other sectoral regulators with mandate to ensure cybersecurity in their specific sectors, in order to enhance much needed coordination.

Indeed, the sole existing coordinating body for information security is the Information Security Management Committee, another GSI body representing of multiple governmental entities, with very limited impact. In its current configuration, this Committee can promote joint regulatory actions and the development and implementation of coordinated policies. However, over the past years, this Committee has been incapable of establishing any concrete initiatives, having promoted no multistakeholder interaction, or proposed not even a single educational, capacity building or compliance-promotion effort.

2.3.2. An "appropriate" way of regulating AI?

It is important to emphasise that both cybersecurity and AI are quintessentially multidimensional matters. Information security is only one for the many dimensions that compose them and, for each dimension them, different regulations, regulators, and regulated entities may already exist.

The success of both cybersecurity an AI governance depend on having a good understanding of how the different component of digital and AI technologies interact, how they are utilised, and what can be the vulnerabilities in their use and deployment.[35]

Sound management of information and infrastructure, good stakeholder coordination, and solid capacity-building are therefore essential. However, as stressed, at the Brazilian level each dimension or component of both AI and cybersecurity is regulated by multiple entities with limited or no coordination at all.

As mentioned in the introduction, Brazil is in the process of elaborating a new AI framework. However, several critiques can be raised as regards both the way in which the framework proposes to regulates cybersecurity aspects of AI and the way in which it proposes to foster coordination amongst sectoral regulators.

In its article 2, the Bill 2338/2023 usefully states that the guarantee of information security and cybersecurity is one of the fundamental principles of AI regulation. However, it subsequently includes a considerable amount of vaguely worded cybersecurity provisions. These include the obligations to adopt "**appropriate** information security measures along the entire AI system lifecycle" (article 17); "perform test to assess the **appropriate** levels of reliability, consistent performance, safety" of AI systems (article 18); or conceive and develop AI systems to achieve "**appropriate** levels of performance predictability, interpretability, correctability, security and cybersecurity assessed through **appropriate** methods" (article 32). [emphasis added]

Appropriate and adequate, along with reasonable, are every lawyer's favourite adjectives, as they can mean literally anything. These flexibility clauses are very welcome to create an agile regulation that does not stifle innovation. However, without an effective mechanism to specify these qualifiers through

[35] M.F. Safitra; M. Lubis; H. Fakhrurroja. Counterattacking cyber threats: a framework for the future of cybersecurity. Sustainability. (2023). https://doi.org/10.3390/su151813369

technical standards[36] or administrative regulation, flexibility turns into legal uncertainty. The opposite of what regulation should bring.

The specification of these elements will require considerable technical skills and is key for the functioning of the AI framework. It is not a coincidence that the European AI Act delegates the specification of such technical, yet vital issues, to standardisation bodies[37], a solution that has raised concerns from human rights advocates[38], but is completely understandable considering the level of technicality that the standardisation of such issues require.

To solve the implementation issue, the Bill proposes to establish a AI Governance and Regulation System, where all sectoral regulators should come together under the leadership of the ANPD. The idea of a coordination system is promising, but the Bill fails to define how it will function in practice. Particularly, it seems a risky gamble to entrust the leadership of the system to the ANPD, considering that is a severely overstretched organ that barely manages to cope with fulfilling its current mission.

Although AI regulation needs to deal with much more than data related-risks, it is understandable that the ANPD is looked to as the leader of such a system. However, to think that ANPD, in its current structure, can effectively lead a new system of such relevance seems overly optimistic. The structure of the Authority should be substantially reformed to have even a minimal chance to successfully coordinate the new AI system.

CONCLUSIONS

As exposed, the relationship between AI and cybersecurity unleashes significant and transformative developments. While it has empowered malicious actors to conduct more effective, far-reaching, and precise attacks, it has also underscored the importance of proactive and adaptive cybersecurity strategies. Indeed, the integration of AI into cyber offensive and defensive capabilities demands a fundamental shift in cybersecurity strategies.

In this context, fostering collaboration between government entities, private sector organisations, and research institutions, becomes essential for Brazil – or any other state – to address the challenges posed by AI in the cybersecurity domain. The adoption of a multistakeholder approach is essential to understand the cyberthreats scenario, develop effective regulations, standards, and governance mechanisms. Indeed, these elements are key to implement robust cybersecurity measures, and

[36] The ISO 27000 and 31000 families of standard is particularly relevant in this regard. For a general overview of existing and under development relevant standards on cybersecurity and AI, see ENISA. Cybersecurity and AI Standardisation. (March 2023).
https://www.enisa.europa.eu/publications/cybersecurity-of-ai-and-standardisation

[37] According to recital 61 of the proposed AI Act "Standardisation should play a key role to provide technical solutions to providers to ensure compliance with this Regulation. Compliance with harmonised standards as defined in Regulation (EU) No 1025/2012 of the European Parliament and of the Council should be a means for providers to demonstrate conformity with the requirements of this Regulation. However, the Commission could adopt common technical specifications in areas where no harmonised standards exist or where they are insufficient." In this respect, in December 2022, the EU Commission adopted the "Draft standardisation request to the European Standardisation Organisations in support of safe and trustworthy artificial intelligence."
See https://ec.europa.eu/docsroom/documents/52376?locale=en

[38] Ada Lovelace Institute. Inclusive AI governance. Civil society participation in standards development. Discussion paper. (March 2023)

promote innovation in defensive AI technologies to safeguard the nation's critical infrastructure and protect its citizens from AI-driven cyberattacks.

However, given the considerations presented in the preceding sections, the current Brazilian institutional arrangement does not seem be fit to provide an effective governance system able to cope with existing cyberthreats, despite the relevant advancements of the country over the most recent years. It seems particularly important that a multistakeholder approach is enshrined in the future strategic and institutional approach adopted by Brazil, not only to increase the quality of policymaking and support it with well-crafted standardisation but, chiefly, to increase the inter-stakeholder coordination and implementation of cybersecurity measures.

Concretely, multistakeholder cooperation should be designed through the development of a "Brazilian Cybersecurity and Digital Transformation System", aimed at facilitating communication, cooperation and – ideally – coordination amongst all governmental entities with these issues. This system should be moulded on the successful experiences of the Brazilian and National Consumer Protection System and the Brazilian Military Cyberdefence System[39]. Ideally such system should be couple with a National Cybersecurity Network facilitating the participation all stakeholders and both the System and the Network should be headed by a much needed National Cybersecurity Agency, able to act as a focal point for cybersecurity governance and regulation.[40]

Hopefully, the aforementioned recommendations will be enshrined in the upcoming proposals on these matters to be issued by the Brazilian Presidency' CNCiber.

[39] National Consumer Protection System (Sistema Nacional de Defesa do Consumidor – SNDC) https://www.consumidor.gov.br/pages/conteudo/publico/6 ; Ordinance No 3.781/GM-MD. (17 November 2020). Military Cyber Defense System (SMDC) https://www.in.gov.br/web/dou/-/portaria-n-3.781/gm-md-de-17-de-novembro-de-2020-289248860

[40] L. Belli *et al*. (2023). *Cit. supra*. N (8).

3. THE LAW ON ARTIFICIAL INTELLIGENCE (AI) IN SOUTH AFRICA IN THE EVOLVING AFRICAN LEGAL LANDSCAPE

PROFESSOR SIZWE SNAIL KA MTUZE, MASEGO MORIGE AND MBALI NZIMANDE

Abstract: This research article will look at the status of AI Laws and Policies in South Africa, the South African Draft AI Strategy (SADAIS) and critique thereof, South Africa's National Artificial Intelligence Policy Framework (NAIPF) as well as African initiatives to regulate Artificial Intelligence as it evolves. The article concludes with a thought on how South Africa is dealing with Artificial Intelligence and the scramble to publish Policy Frameworks to govern it.
Keywords: African Intelligence (AI), South Africa, AI Strategy, AI Policy, African Framework on AI.

INTRODUCTION

The regulation of Artificial Intelligence in South Africa has been a contentious and ongoing issue discussed by various African Scholars. (Adeyoju: 2018:1, Ncube, et al:2023 & Snail & Morige:2024). This research article will look at the status of AI Laws and Policies in South Africa, the South African Draft AI strategy and critique thereof, South Africa's National Artificial Intelligence Policy Framework as well as African initiatives to regulate Artificial Intelligence. South Africa does not have a formal AI Policy and it is because of this that mention has been made to the various pieces of legislation. These pieces of legislation are currently being used to govern AI in the absence of the policy. This article will then conclude with a thought on how South Africa is navigating the evolution of AI.

3.1. THE STATUS OF AI IN SOUTH AFRICAN LAW

AI has had an impact on a variety of industries in the modern world and the legal profession is no exception. The 4th Industrial Revolution has had a negligible effect on the legal profession and this is as a result of the immunity that it enjoys when compared to other professions. This immunity is protected by professional rules, guidelines and ethics. (Adeyoju: 2018: 2-3) However, it seems that this immunity will not be functional for much longer because a majority of the protections are being eroded as the laws on AI evolve (Adeyoju: 2018: 2-3). AI has percolated into the profession and the need has arisen for there be legislation that will regulate and guide its ethical use. AI has emphatically made its presence known and there has been an escalation in the need for it to be increased rapidly.

The issue we face in South Africa is that there is currently very little existing legislation, regulatory mechanisms or policies that will do this. (Adams: 2021:13 & Brand: 2022:142). For those who argue that there is such existing legislation, they have neglected to note that it may only be applied in a general sense and that its relevance to AI is limited. What is lacking is specific regulatory frameworks and policies governing how we use AI in our country. (Adeyoju: 2018:2-3).

3.1.1. PC4IR REPORT

Following the development of AI and the lack of legislation designated for AI-related matters, the South African President in 2019, initiated the Presidential Commission on Fourth Industrial Revolution Commission[41] (which then issued the Presidential Commission on Fourth Industrial Revolution

[41] Department of Telecommunications and Postal Services, Terms of Reference for the Presidential Commission on the Fourth Industrial Revolution GN 209 in GG 42388 of 2019- 04-09 (https://www.gov.za/sites/default/files/gcis_document/201904/42388gen209.pdf)

Commission (PC4IR Report) [42] which came up with eight key recommendations, including the establishment of an artificial intelligence (AI) institute and the review and amendment (or creation) of policy and legislation. The PC4IR Report put forward key points which are the pillars of AI's development in South Africa. (SAAIP: 2024:21). According to the South African government setting up of the AI Institute is a summary of all of the intended actions of the government in ensuring the smooth transition of AI into both the public and private sector and how it will enhance the already existing skills and research.

3.2. SOUTH AFRICA'S DRAFT AI STRATEGY & CRITIQUE

The South African government has put together a discussion document named the South Africa's Draft AI Strategy (SADAIS) (SAAIP: 2024:25). It discusses its priorities, intentions and objectives for the adoption of AI into South Africa's various sectors and to bring about the envisioned economic advancement (SAAIP:2024:3). The discussion document is divided into 3 (three) sections and each section touches on a different aspect. The plan by the government is to facilitate a better use of AI in the future through a variety of measures. These measures are the,

> *"creation of policy and regulatory experiments; set of positive goals for what South African society require from AI; building an understanding of the AI technological possibilities; management of negative AI impacts on society and industry and providing certainty to society on this rapidly evolving AI technology through flexibility and accommodation of skills, software, innovations and applications". (SAAIP: 2024:8)*.

In order for the SADAIS to succeed, there are 8 (eight) pillars on which it will rely. These pillars are envisaged to ensure that all sectors are accounted for in this transitory period. The most important pillar is the one that speaks to the need for there to be separate legislation which will highlight that AI is important and that its field of technology is equally important (SAAIP: 2024:8). Another equally important pillar is one that touches on the belief that South Africa truly has the potential to be valuable and bring about positive change (SAAIP: 2024:22).

With regard to the actual adoption of AI, the PC4IR report states the terms and conditions of the approach which must be taken (SAAIP:2024:15). The approach must be one that is inclusive, integrated, adaptive and mindful of the socio-economic impact (SAAIP:2024:24). Those who find themselves tasked with regulating should concentrate on how they are going to overcome the hurdle of the lack of exploration of AI in laws and regulation (SAAIP: 2024:24). As a result, those tasked with making policy should focus on building trust among people in AI-driven systems. This trust can be built through the development of intelligible frameworks and clear attributions of accountability.

The SADAIS consists of 4 (four) phases and they span from the year 2023 to 2026. Phase 0 (zero) is scheduled for 2023 and the plan is that strategy formulation takes place and strategies are developed (SAAIP: 2024:29). Phase 1 (one) is scheduled for 2024 and the plan is to activate initiatives, test them and assess their effectiveness and efficiency (SAAIP:2024:23).

Phase 2 (two) is scheduled for 2025 and the plan is to expand execution through the activation of more institutes in order to achieve strategic objectives. Phase 3 (three) is scheduled for 2026 and is the

[42] Commission on the Fourth Industrial Revolution , Summary Report & Recommendations GN 591 in GG 43834 of 2020-10-23 (https://www.gov.za/sites/default/ files/gcis_document/202010/43834gen591.pdf)

finale where all the existing initiatives are accelerated to a national level (SAAIP: 2024:23). Technology can be used as a tool of choice and it will have an impact on two sectors, namely the social and economic sectors. It aims to achieve the following 4 (four) outcomes: AI Predictive maintenance abilities, AI Logistics optimization and Automated services, AI Diagnostic abilities and AI Analytical abilities (SAAIP:2024:23).

AI also has benefits that will benefit the country and accelerate the transition into the new normal of a AI-driven modern country using the four-phase (strategy formulation, activate initiatives, expand execution and accelerate execution) plan (SAAIP:2024:35).

3.2.1. Critique: The Good

The SADAIS was published by the Department of Communications and Digital Technologies (DCDT) in April of 2024 at a time when Africa is undergoing a comprehensive review of AI policies and laws. The purpose of the document was to commence talks and strategizing between the public and private sector. (SAAIP: 2024:35). These talks were initiated with the aim of facilitating AI innovation, government-led AI initiatives, regulatory frameworks and principles and ultimately, the development of a national AI policy (Bhagattjee:2024). It was a working paper and it was a step in the right direction with regard to the regulation of AI. (Bhagattjee:2024) If it happens that it is adopted as a White paper, it will prove useful as it is well compiled and would play a key role as a regulatory and governance tool. The SADAIS provided key proposals and insights into the government's approach. One of the key proposals was to ensure that any ethical considerations relating to AI are addressed appropriately under the legal framework to guard against any potential harm as an important component of the SADAIS is that it considers that the future use of AI could cause harm to humans and raise ethical concerns. (Bhagattjee:2024)

As a result, this requires regulation on aspects such as the social risk of loss of employment, dangerous outcomes which would come with increased criminal behaviour, the risks that come with robotic or autonomous devices that are AI-centric and the risks posed by the potential detriment humanity faces from AI. The SADAIS advocated for AI literacy as it is needed by South Africa and it can be provided through education and training and by investing in technology start-ups (Bhagattjee:2024). The hope is that this Discussion Document is reworked and that it is then published with input from key stakeholders from both the private and public sector as well as the AI Expert Advisory Council and any other relevant AI bodies (Bhagattjee:2024).

When the document was launched, the Minister of DCDT alluded to the type of approach that government would take in its approach to regulating AI as it is not set out clearly in the document (Bhagattjee:2024). If one were to look for the most positive take-away from this document, it is that it takes into account how different jurisdictions around the world regulate AI and implement it using effective mechanisms to foster and encourage AI use and development. This is done whilst also striking a balance between risk-management, assessment of harms and a consideration of major and minor ethical risks as they are equally important (Bhagattjee:2024).

3.2.2. Critique: The Bad and Ugly

The SADAIS was a 53-page long document and it has a disclaimer which states that it is a discussion document. (Pierce:2024). Pierce is in agreement with the disclaimer and further voices that it lacks clear deliverables and that is complicated and not up to standard. The Plan is lengthy, it contains a high

volume of jargon and has a number of unfinished thoughts (Pierce:2024). In November of 2022, the Artificial Intelligence Institute of South Africa was set up and things seem to be progressing slowly as even its website has not yet been updated since March 2023 (Pierce:2024).

The issue that the slow progression poses is that the rollout of the AI plan is centred around this Institute and this could seriously delay things. Another critique is the unrealistic timetable that has been set for the adoption of AI (Pierce:2024). Other countries have given themselves much more time to adopt it whereas South Africa has set themselves 12-month deadlines to achieve the impossible (Pierce:2024). Pierce makes mention of Rwanda's National AI Policy and in comparison to South Africa's, he is of the opinion that it is a far more practical document that is uncluttered and has very little room for misinterpretation.

Therefore, Pierce recommendation was that the entire SADAIS is reworked and that it is released timeously as South Africa runs the risk of getting left behind while the rest of the world passes AI Acts and publishes practical policies (Pierce:2024) . Seth Thorne has also given his views on the document and he shares sentiments which are similar to those of Pierce. Thorne shares an important view with Pierce which is that the draft touches on Data Sovereignty which will be managing the data that will be needed for the AI training however, it fails to address the challenges that will come about in securing the computing power necessary for the achievement of its objectives (Thorne:2024).

3.3. South Africa: National Artificial Intelligence Policy Framework

The South African National Artificial Intelligence Policy Framework (NAIPF) was drafted in August of 2024 and it can be considered to be the first step in the actual development of a National AI Policy. It follows the Draft AI Strategy Document (SAAIP: 2024:1). The NAIPF is intended to serve as the foundational basis for creating AI regulations and potentially an AI Act in South Africa, and guide the development of robust regulatory mechanisms that ensure that AI applications are safe, ethical and in the public interest. (Rosenburg & Madondo: 2024:1). The rationale for the development of an AI policy document in South Africa was that it is imperative that there is a set of guidelines to ensure the responsible and ethical use of AI across all societal sectors. The rapid advancement of AI technologies offers opportunities for an enhanced quality of life, economic advancement and an improvement in public services (Rosenburg & Madondo:2024:3).

However, these opportunities are at risk of never being actualized due to the risks that are posed by letting AI develop without any policy to regulate it. Thus, having an AI policy will provide the foundation for AI to be regulated and for there to eventually be an AI Act (Rosenburg & Madondo:2024:3).

The NAIPF acknowledges global trends in AI governance and the need to harmonise with international standards, pushing South Africa to develop its own AI policies. It seeks to align with international norms and standards to ensure ethical and effective AI deployment. (Rosenburg & Madondo: 2024:5).The NAIPF has 12 (twelve) fundamental components which can be characterized as the support behind the implementation of the goals and objectives of the National AI Policy (Rosenburg & Madondo:2024:5). Talent and capacity development is one of the components and its aim is to ensure that South Africa has a robust AI talent pool. Digital infrastructure is also one and its aim is to create an environment which will foster AI innovation. Research, development and innovation is component number three and its aim is to be the driving force behind AI innovation and ensure the advancement of

technological capabilities. Component number four is Public Sector Implementation and its aim is to use AI to enhance the efficiency of our government (SANAIPF:9).

Component number five is Ethical AI Guideline Development and they are there to ensure that the use of AI is ethical and responsible. Component six requires Privacy and Data Protection has the aim of safeguarding the personal information of all people who will be bound by the National AI policy (SANAIPF:9). Safety and Security is component number seven and its aim is to protect citizens and ensure that cybersecurity protocols are safeguarded by AI systems. Transparency and Explainability have the aim of building trust amongst members of the public in component number eight. If the National AI Policy provides clear and understandable information on AI, it will be easier for the public to understand the AI systems (SANAIPF:10). In order to ensure that AI is deployed equitably there must be Fairness and Mitigating Bias as per component nine (SANAIPF:10).

The importance of the Mitigating Bias is to ensure that it identifies any bias that might be present in the AI systems. Component number ten is Human Control of Technology and it is there to ensure that the AI systems have human oversight and to ensure the prioritisation of a human-cantered approach within the systems. Professional Responsibility is component number eleven and it creates a code of conduct for AI professionals and ensures the upholding of ethics as per component eleven. The final component is the Promotion of Cultural and Human Values and its aim is to ensure that the development of AI is aligned with societal values and promotes environmental sustainability and human well-being (SANAIPF:11). The abovementioned key pillars are crucial to the meaningful contribution of AI technologies to important sectors such as healthcare and education (SANAIPF:12).

The NAIPF outlines key pillars such as robust Data Governance Frameworks, Infrastructure Enhancement, and Significant Investments in Research and Innovation, which the DCDT believes are crucial components to create an enabling environment where AI technologies can thrive and contribute meaningfully to sectors such as healthcare, education and public administration (Rosenburg & Madondo:2024:6) Overall, the NAIFP seeks to lay the groundwork for South Africa to emerge as a leader in AI innovation while addressing challenges and opportunities in a holistic and sustainable manner (Rosenburg & Madondo:2024:6).

3.4. Overview of Regulation of Artificial Intelligence in Africa

AI is slowly making it to the meeting agendas of organisations globally including across Africa. Such that, there is an important piece of African International law namely the African Union Convention on Cyber Security and Personal Data Protection[43]. It has limited AI regulatory properties and spearheads matters of data protection, cybercrime and cyber security in the African continent. Article 9 of the Convention regulates data processing and this is inclusive of the automated processing of personal information by AI. Article 14.5 of same confers rights on all data subjects that they may not be affected by legal effects that significantly affect them solely based on automated data processing (Orji,et al:2024:172).

[43] African Union Convention on Cyber Security and Personal Data Protection (2014) (https://au.int/en/treaties/african-union-convention-cyber-security-and-personal-data-protection)

In the AU Digital Strategy Information for Africa for 2020-2030[44] a proposition was made in Kenya and makes extensive references to the governing of AI. The proposition is that there be a continent-wide digital governance African Peer Review Mechanism on AI use. It would be applicable to member states and it would prescribe rules on AI with a basis on solidarity and to ensure that Africa is cooperative with forthcoming digital infrastructure (Ncube,et al:2023:69). One of the first African countries to establish a national policy and institutional framework to govern AI is Mauritius. Its national AI strategy was established in November of 2018 and its aim is to address any ethical concerns surrounding the development and use of AI as well as to promote capacity building (Orji,et al:2024:69). As far as Africa is concerned as of March 2024. 9 (nine) out of the 54 (fifty four) states had established AI policy frameworks and only 2 (two) had already established institutional framework (Orji,et al:2024:171).

The Continental AI Strategy calls for unified national approaches among AU Member States to navigate the complexities of AI-driven change, aiming to strengthen regional and global cooperation and position Africa as a leader in inclusive and responsible AI development.[45] The AU AI Strategy contains a key action point and that is the building of a AI knowledge base speaking on AI use cases and the monitoring of the implementation of the recommendations of the Strategy (Alayande & Adams:2024:3).

CONCLUSION

What we can conclude from all that has been said is that South Africa does not seem adequately prepared to deal with the multifaceted and evolving AI. There is an evident lack of regulatory policies, undefined laws, critical judgements having been handed down and 53-page discussion documents which do not have clear directives and implementation procedures. It also seems that the South African government has seem the deficiencies in the SADAIS hence it has rushed within months to develop the NAIPF which has been received less critically that the previous SADAIS.

References

Adams, N., 2024 Parker v Forsyth no lessons for using ai for legal-research (https://www.michalsons.com/blog/parker-v-forsyth-no-lessons-for-using-ai-for-legal-research/66884.)

Adeyoju, A. (2018) *Artificial Intelligence and the Future of Law in South Africa*.

Alayande, A., Adams, R., (2024) "Africa Now Has a Continental AI Strategy: What Next?" August

Bhagattjee, P. and Stephens, S (2024) The AI National Policy: South Africa's initial step to establish an AI policy and regulatory framework. (https://www.werkmans.com/legal-updates-and-opinions/the-ai-national-policy-south-africas-initial-step-to-establish-an-ai-policy-and-regulatory-framework/).

Brand (2022) Responsible Artificial Intelligence in Government: Development of a Legal Framework for South Africa" 14(1) in JeDEM

[44] The Digital Transformation Strategy for Africa (2020-2030) (https://au.int/en/documents/20200518/digital-transformation-strategy-africa-2020-2030)
[45] *Ibid*.

Pierce, L. South Africa's Draft AI Plan: Not Good Enough (2024) https://www.linkedin.com/pulse/south-africas-draft-national-ai-plan-good-enough-lucien-pierce-gp5cc.

Rosenburg, W and Madondo, N (2024) The National AI Policy Framework: A step closer to aligning with international trends (https://www.werksmans.com/legal-updates-and-opinions/the-national-ai-policy-framework-a-step-closer-to-aligning-with-international-trends/)

Snail Ka Mtuze, S., Morige, M. (2024) Towards Drafting Artificial Intelligence (AI) Legislation In South Africa in Obiter

Thorne, S. (2024) South Africa's proposed AI plan needs a rework: experts (https://businesstech.co.za/news/government/768147/south-africas-proposed-ai-plan-needs-a-rework-experts/.)

Adams,N (2021) South African Company Law in the Fourth Industrial Revolution: Does Artificial Intelligence Create a Need for Legal Reform? (LLM thesis, Wits University)

Marwala, T. and Mpedi, L.G., 2024. Artificial Intelligence and the Law. (Palgrave)

Ncube,C., Oriakhogba,D. Rutenberg, Schonwetter, T. (2023) Artificial Intelligence and the Law in Africa (Lexis Nexis)

Orji, U. Regionalising the Governance of AI in Andersen, L.H., Broeders, D. and Csernatoni, R., (2024). "Emerging and disruptive digital technologies: National, regional, and global perspectives".

African Union (2024) Continental Artificial Intelligence Strategy (https://au.int/en/documents/20240809/continental-artificial-intelligence-strategy)

African Union Continental Artificial Intelligence Strategy (https://au.int/sites/default/files/documents/44004-doc-EN _Continental_AI_Strategy_July_2024.pdf)

Commission on the Fourth Industrial Revolution, Summary Report & Recommendations GN 591 in GG 43834 of 2020-10-23 (https://www.gov.za/sites/default/files/gcis_document/202010/43834gen591.pdf) Department of Telecommunications and Postal Services, Terms of Reference for the Presidential Commission on the Fourth Industrial Revolution GN 209 in GG 42388 of 2019-04-09 (https://www.gov.za/sites/default/files/gcis_document/201904/42388gen209.pdf)

DCDT, South Africa National Artificial Intelligence Policy Framework (https://www.policyvault.africa/policy/south-africa-national-artificial-intelligence-ai-policy-framework-2024/)Electronic Communication and Transactions Act 25 of 2002.

Smart Africa (2021) Artificial Intelligence in Africa (https://smartafrica.org/knowledge/artificial-intelligence-for-africa/)

South Africa's Artificial Intelligence (AI) Planning: Adoption of AI By The Government (https://www.dcdt.gov.za/images/phocadownload/AI_Government_Summit/National_AI_Government_Summit_Discussion_Document.pdf).

The Digital Transformation Strategy for Africa (2020-2030)
(https://au.int/en/documents/20200518/digital-transformation-strategy-africa-2020-2030)

4. BUILDING SMART COURTS TROUGH LARGE LEGAL LANGUAGE MODELS? EXPERIENCE FROM CHINA

ZIJING LIU, SHAOYU LIU AND YIN LIN

Abstract. Artificial intelligence is already all around us and has been applied to almost every aspect of society and business. One of the most striking innovations in the application of AI has been the introduction of large legal language models in judicial decision-making. There has been growing interest in the use of AI in legal systems worldwide in recent years, particularly in the role of judges. The main question addressed in this Article is that what should be the potential legitimacy, weaknesses, and limitations of large legal language models in judicial scenery. To address it, it takes the Chinese smart court construction as an example and studies 133 cases of smart courts from 2017 to 2024. It summarizes four patterns from Shanghai city, Zhejiang province, Jiangsu province, and Shenzhen city. Based on this, this article analyzes the achievements and shortcomings of the application of large language models in China's smart courts.
Keywords. Artificial intelligence, large legal language models, smart courts, Judicial decision-making, empirical study.

INTRODUCTION

Technological innovations such as big data, cloud computing, and artificial intelligence have created a worldwide digital revolution regarded as 'the Fourth Industrial Revolution' which impacts everyone's life (Klaus Schwab, 2016). Already, artificial intelligence is all around us and has been applied to almost every aspect of society and business, from assigning credit scores to assessing the criminal risk of people (Antunes H. S. et al., 2024, p.281). One of the most striking innovations in the application of AI in the justice system in recent years has been the introduction of large legal language models in judicial decision-making and other assistance jobs (Bin Wei, 2024).

In recent years, there has been growing interest in the use of AI in legal systems, particularly in the role of judges (Ulenaers, 2020). In 2023, a Colombian judge used the AI chatbot ChatGPT in preparing a ruling in a children's medical rights case by asking the chatbot whether an autistic child's medical insurance should cover the cost of related therapies (Luke Taylor, 2023). Later this year, an intellectual property law Judge in England used ChatGPT to assist judicial decision-making, such as summarizing information on the law in a particular field (Gareth Corfield, 2023). Compared to Colombia and England, East Asian countries such as India and China use the large language model in the judiciary sector more aggressively. The Indian Supreme Court has set up an AI Committee with a focus on the translation of legal documents; process automation; increasing administrative effectiveness; automating forecasting, prediction, and filing; scheduling of cases; and early case resolution using chatbots (Gandhi & Talwar, 2023). The Chinese government is even more ambitious. It issued a '*New Generation Artificial Intelligence Development Plan*' in July 2017, advocating to establishment of an AI-powered 'Smart Court'.[46] In Oct 2024, Zhang Jun, president of the Supreme People's Court, stressed the need to explore the use of artificial intelligence technology to empower the judiciary and promote the deep integration of artificial intelligence and judicial work (Zhang Jun 2024).

[46] State of Council of People's Republic China, *Notice of the State Council on Issuing a New Generation Artificial Intelligence Development Plan*, 2017, https://www.gov.cn/zhengce/zhengceku/2017-07/20/content_5211996.htm.

The implementation of large legal language models in judgment is controversial. Despite the potential advantages of robot judges, it raises significant concerns, such as the concerns of algorithm bias, non-transparency, inaccuracy, weak interpretability, hallucinations, lack of human empathy, data privacy, and security problems (Magnus Kristoffersson, 2024). Consequently, scholars around the world have highlighted that the application of generative AI, particularly the large legal language model, cannot be a substitute for human judges (Parikh et al., 2023). The main question addressed in this Article is, thus, what should be the potential legitimacy, weaknesses, and limitations of large legal language models in judicial scenery. To address it, it takes the Chinese smart court construction as an example and studies 133 cases of smart courts from 2017 to 2024. It summarizes four patterns from Shanghai city, Zhejiang province, Jiangsu province, and Shenzhen city. Based on this, this article analyzes the achievements and shortcomings of the application of large legal language models in China's smart courts.

4.1. METHODS

4.1.1. LEGAL EMPIRICAL ANALYSIS

Data analysis

This article collects 133 case samples of smart court construction from 55 regions in China from 2017 to 2024. It analyses specific cases utilizing the large legal language model and discusses the experiences, achievements, and shortcomings of China's foundation model construction. The primary textual source is '*the China Court Informatization Development Evaluation Report*' conducted by the Chinese Academy of Social Sciences. This report is published annually since 2017, making it the most authoritative and systematic public resource for the construction of smart courts in China (Tian He, 2024).

Face-to-face interviews

In addition, this article also uses questionnaires and interviews to conduct in-depth interviews with judges who use AI large language models to understand the current status and potential problems of the smart court.

Normative analysis

Normative analysis method is a unique method of jurisprudence. It mainly focuses on the legality of law, the operation effect of law, the substantive content of law, and examines the constituent elements of law in an all-round way.

4.2. DISCUSSION

4.2.1. BACKGROUND: THE MOTIVATION OF CHINA'S SMART COURT CONSTRUCTION

Through the holistic approach of 'Smart Court' construction, China has significantly advanced the application of the foundation model in the field of adjudication, which is closely related to the functional requirements and inherent challenges faced by courts: Firstly, the contradiction between the increasing caseload and limited judicial personnel has intensified, leading to inefficiencies in the judicial process. In 2015, to address the issue of 'difficulty in filing cases,' Chinese courts initiated reforms to the case registration system, resulting in a surge of disputes entering the courts and placing immense pressure on their adjudication capacity. In 2022, the average number of cases concluded per judge in grassroots courts reached 274, with some exceeding 400, yet the number of judges nationwide

has not increased significantly over the past decade, further exacerbating the case-to-judge imbalance. Secondly, judicial fairness needs improvement. Despite hierarchical trial supervision mechanisms such as second-instance final judgments and retrials, inadequate case quality inspection mechanisms have led to repeated instances of inconsistent judgments for similar cases and misjudgements. Thirdly, judicial credibility remains insufficient, with 'visible justice' not fully achieved, and public trust in the judiciary requires further enhancement (Jia Yu, 2024).

4.2.2. Four patterns: utilization of large legal language models in Chinese smart court

The construction of smart courts encompasses a comprehensive process that integrates informatization, datafication, and intelligence across various stages such as case filing, trial, supervision, and management. Among these, the trial phase prominently showcases the application and technological characteristics of the foundation model, specifically including (Wei Bin, 2022):

- Similar Case Recommendation. It involves retrieving and presenting similar cases and their corresponding judgments based on the current case being heard.
- Legal Judgment Prediction. This entails extracting key information from judgments, categorizing it, and utilizing text classification techniques to forecast the outcome of the current case, including charges, applicable laws, and sentences.
- Automated Generation of Legal Documents. It involves constructing a knowledge graph of the case based on trial data and legal knowledge and then employing machine learning and natural language generation techniques to automate the generation and proofreading of legal documents.

Currently, the development of foundation models is primarily driven by local pilot projects, resulting in four distinct patterns as follows.

4.2.2.1. Shanghai Model

In 2017, the Shanghai Higher People's Court introduced the 'Intelligent Assistant System for Criminal Cases' (206 System), which leverages AI for evidence analysis, unifying evidence standards, formulating evidence rules, and constructing evidence models. This aims to achieve the judicial goals of uniform law application and prevention of miscarriages of justice. The 206 System employs new AI technologies such as optical character recognition, natural language processing, intelligent speech recognition, element extraction, and machine learning to provide guidance for case handlers in collecting and fixing evidence, enabling judgment, verification, control, and supervision of evidence. Through this system, flaws and contradictions in evidence can be promptly identified and flagged for case handlers, thereby preventing miscarriages of justice (Cui Yadong, 2020).

4.2.2.2. Zhejiang Model

Led by the Zhejiang Higher People's Court, the Full-process Intelligent Trial System (FITS) 'Xiaozhi' was developed, capable of tasks like legal information extraction, evidence classification, question generation, dialogue summarization, judgment prediction, and judgment document generation (Yu Shujun, 2019). The system first extracts elements from legal texts to assist judges in effectively identifying the essence of cases. It then verifies the consistency of all evidence to demonstrate its validity. Additionally, it features an automatic questioning robot that assists judges in posing questions during trials, both procedural and factual. The system can also summarize points of contention during

court debates under a multi-task learning framework, generating real-time trial records automatically. Lastly, it proposes a natural language generation method based on attention and counterfactual reasoning to produce court judgments. Currently, the system can assist judges in handling specific types of simple cases such as financial loan contracts, private lending, motor vehicle accidents, theft, and divorce, enhancing case handling efficiency.

4.2.2.3. Suzhou Model

In Jiangsu, the Suzhou Intermediate People's Court has also piloted a generative AI-assisted case-handling system. Building upon electronic case file data and legal knowledge data accumulated from previous paperless case-handling initiatives, the court has integrated the 'General AI foundation model' technology to create a specialized prophecy model tailored for courts, boasting multi-modal document comprehension, legal semantic cognition, and natural language interaction capabilities. This AI-powered system can accurately identify and present factual elements required by judges within electronic case files, including their original sources. Its built-in annotation and element backfilling functions facilitate judges in reviewing case files, retrieving evidence, and organizing facts. Furthermore, the system can mimic judicial thinking to organize language and generate relevant legal documents, with accuracy rates exceeding 95% for party information and 'fact finding' sections, and around 70% completeness for reference 'judgments' (Suzhou Intermediate People's Court, 2023).

4.2.2.4. Shenzhen Model

On June 28, 2024, the Shenzhen Intermediate People's Court launched an AI-assisted trial system that supports judges throughout 28 critical nodes and 57 auxiliary nodes, from case filing to closure. During case review, the system enables precise data tracing and comparison, facilitating refined and user-friendly information processing. During trials, it real-time assists in evidence verification and logical review, enhancing trial quality and efficiency. During judgment, it matches similar cases, applicable laws, and authoritative viewpoints to ensure uniformity in judgment standards. Additionally, it innovatively employs a foundation model tree-structured prompt engineering component to manage judgment standards. Lastly, the system incorporates a self-learning and feedback mechanism, dynamically optimizing based on judges' usage and actual judgment outcomes (Guangdong Higher People's Court, 2024).

4.2.3. Achievements and Problems

Although China has made certain achievements in the development of law and artificial intelligence, there are still numerous issues, mainly manifested in three aspects: technical issues, issues of justice, and institutional issues.

4.2.3.1. Technical Concerns

The large legal language model has exposed problems such as weak interpretability and the generation of false content due to 'hallucinations' in the judicial field. Firstly, the foundation model's use of neural network algorithms leads to the 'black box' problem in algorithmic decision-making, thereby rendering the process and results of legal predictions lacking in transparency and interpretability (Wei Bin, 2024 b). Secondly, foundation models still suffer from data 'hallucinations' that may compromise the accuracy of results. Thirdly, judicial artificial intelligence systems are still primarily expert systems, and constructing expert graphs requires extensive manual annotation and organization, which might paradoxically increase judges' workload. Currently, China's foundational model is still in its infancy, with notable flaws that prevent it from replacing legal professionals and relegating it to an auxiliary role.

In tasks such as legal prediction, the foundation model still struggles to handle the core work of legal professionals, including legal reasoning, legal argumentation, judicial proof, legal interpretation, and judgment of complex cases.

4.2.3.2. Justice and Ethics Concerns

The existence of trial assistance systems can easily make judges susceptible to flawed preconceived judgments, leading to psychological anchoring effects and potentially even being 'monitored' and 'hijacked' by artificial intelligence, thereby affecting judges' discretion. Moreover, if paperless, visualized, and integrated artificial intelligence systems are fully implemented in courts of different levels in the future, they might undermine judicial independence.

4.2.3.3. Institutional concerns

Judicial artificial intelligence is significantly constrained by local fiscal capacity and regional economic development levels, resulting in significant disparities in development between regions. Taking the construction of the Guangzhou Internet Court as an example, just the first phase requires a budget of 15.59 million yuan (Zhou Xiang, 2021). Additionally, the development of judicial artificial intelligence systems cannot be achieved without the support of technology companies. Regions like Beijing, Hangzhou, Shenzhen, and Shanghai enjoy distinct geographical advantages, which will lead to a fragmented market landscape.

CONCLUSION

This article investigated the use of large legal language model in China's smart court construction as it exists today regarding their skill to solve legal problems. The conclusion based in this is that large legal language model such as ChatGPT or other AI chatbot can be used as robot judges in the judicial decision-making, and that the future is already coming. The large legal language model is used not only in the judicial decision-making, but also in the public legal service, which needs a further discussion (Dai Xin, 2024). Besides, the 'robot lawyer' as well as 'robot judge' calls for more empirical studies.

References

Schwab, K. (2016, Jan). *The Fourth Industrial Revolution: What It Means and How to Respond*, World Economic Forum , Website. https://www.weforum.org/agenda/2016/01/the-fourth-industrial-revolution-what-it-means-and-how-to-respond/.

Antunes, H. S. et al. (2024). *Multidisciplinary Perspective on Artificial Intelligence and the Law*. Springer. https://doi.org/10.1007/978-3-031-41264-6.

Wei, B. (2024). Judicial Application and Specification of Large Language Model, *Oriental Law*, vol.5, pp.57-73. DOI:10.19404/j.cnki.dffx.2024.05.012.

Ulenaers, J. (2020). The Impact of Artificial Intelligence on the Right to a Fair Trial: Towards a Robot Judge? *Asian Journal of Law and Economics*, 11(2). p. 2, https:// doi.org/10.1515/ajle-2020-0008.

Taylor, L. (2023, Feb. 2), *Colombian Judge Uses ChatGPT in Ruling on Child's Medical Rights Case*, CBS News. https://www.cbsnews.com/news/colombian-judge-uses-chatgpt-in-ruling-on-childs-medical-rights-case/.

Corfield, G. (2023, Sept. 14). *British judge uses 'jolly useful' ChatGPT to write ruling*. Telegraph.https://www.telegraph.co.uk/business/2023/09/14/british-judge-uses-jolly-useful-chatgpt-to-write-ruling/.

Gandhi, P. & Talwar, V. (2023). Artificial Intelligence and ChatGPT in the Legal Context, *Indian Journal of Medical Sciences*, vol. 75. pp. 1–2. doi:10.25259/IJMS_34_2023.

Kristoffersson, M. (2024). The Concept of Robot Judges Using Generative Artificial Intelligence and the Rule of Law. In: Rigmor Argren (ed.), *Rule of Law in a Transitional Spectrum*, pp. 369-388.

Parikh PM, Shah DM, Parikh KP. (2023). Garcia JJ. ChatGPT and a controversial medicolegal milestone. *Indian Journal of Medical Sciences*, 75(1).

Tian He.(2024). China 'Court Informatization Development Report (No.8, 2024), *Social Sciences Academic Press*.

Jia Y, (2024), On Digital Courts, *Chinese Journal of Law*, vol.46, No.4, pp.3-20.

Wei B, (2022). Legal Argumentation Analysis of the Interpretability Challenge in Judicial Artificial Intelligence. *Legal System and Social Development*, vol. 30, No.4, pp.76-92.

Cui Yadong. (2020). Application and Governance of Artificial Intelligence, *Reform of Public Administration*, vol.42, No.6.

Yu Shujun (2019, Sept. 26), *Zhejiang Court first 'Phoenix Financial smart court'*, http://zjnews.china.com.cn/yuanchuan/2019-09-24/189935.html.

Suzhou Intermediate People's Court. (2023, Nov 23). The 'Generative AI-Assisted Case Handling System' Approved for Provincial Court Pilot, *Jiangsu Legal Daily*, http://www.zjrmfy.suzhou.gov.cn/fypage/toContentPage/xwzx/82a07a488c3068a6018c37980cca0012.

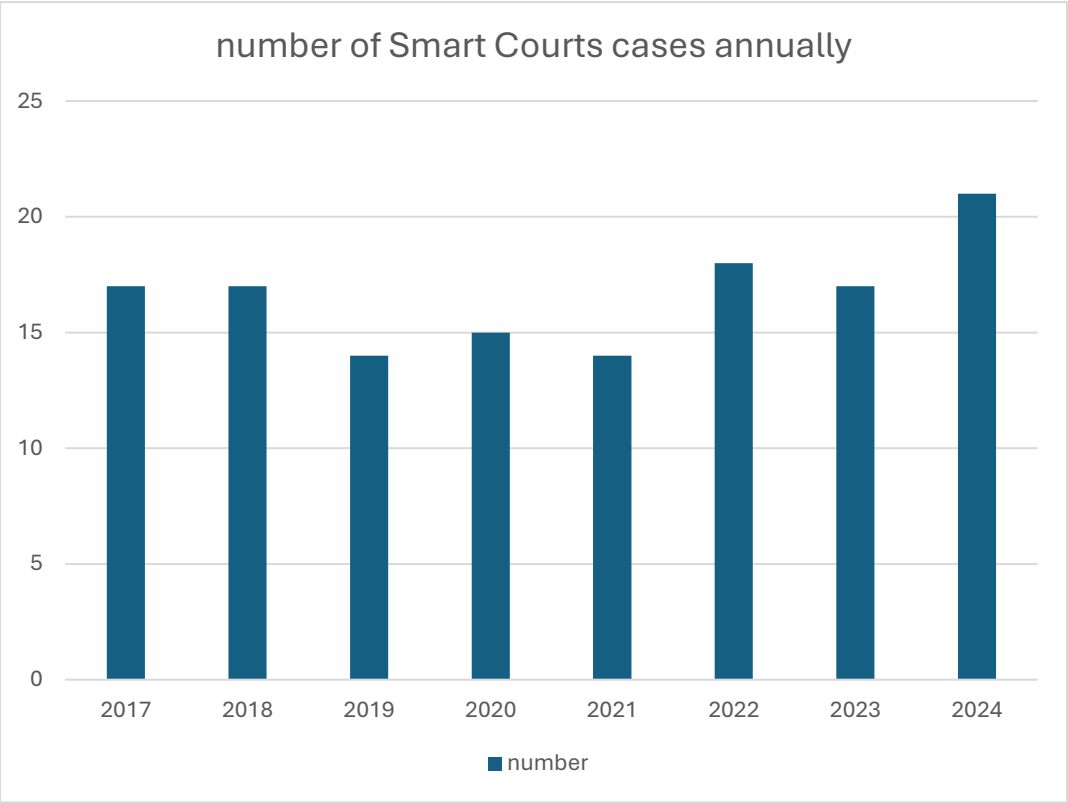
Guangdong Higher People's Court.(2024, June 7). Shenzhen Intermediate People's Court's AI-Assisted Judicial System Officially Launched, https://news.southcn.com/node_d16fadb650/3eb7350386.shtml.

Wei B., (2024 b). Legal Argumentation Analysis of the Interpretability Challenge in Judicial Artificial Intelligence. *Legal System and Social Development*, vol. 30, no.4, pp.76-92.

Zhou X., (2021) The Formation Mechanism and Future Development Trends of Smart Courts. *Journal of Xi'an Jiaotong University (Social Sciences)*, vol. 4, no.3, pp.131-140.

Dai X., Who Wants a Robo-Lawyer Now?: On AI Chatbots in China's Public Legal Services Sector, *Yale Journal of Law & Technology*. Volume 26, Issue 3.pp.528-559.

Figure 1. Number of smart court cases annually



Note. Chinese Academy of Social Sciences makes reports concerning the smart court annually. The above table is the number of smart court samples each year from 2017 to 2024.

5. Fox Guarding the chickens – Bias in Risk Management Obligations for high-risk AI Systems under the EU AI Act

Nils Brinker and Richard Skalt

Abstract. In the context of the regulation of so-called high-risk AI applications by the EU AI Act, the obligation to conduct risk management plays a decisive role. In theory, manufacturers and operators of these systems must already mitigate the risks posed by their systems during the development phase. However, this paper argues that there is a fundamental bias on the part of manufacturers and operators, which threatens to result in third-party risks in particular not being adequately taken into account. It is also shown that the concretization mechanisms of the relatively abstractly formulated AIA play a critical role in ensuring that third-party risks receive appropriate attention.
Keywords: EU AI-Act, Risk Management, Principal Agent Relationship, Principle based Regulation

Introduction

The European AI Act[47] has taken on the task of creating a regulation for AI as a technology that is still in development. For the category of "high-risk" AI systems in particular, a product safety regulation has been chosen that permits the development and marketing of such systems provided that certain requirements are met. A key aspect here is the implementation of risk management, which in theory should reduce the risks of these systems to an acceptable level. Since this risk management must be carried out by the manufacturer of the AI systems, this paper will address the question of whether such a methodology adequately takes into account the risks to third parties, i.e. to persons who are not the manufacturers, users or operators of these systems.

5.1. Discussion

5.1.1. The EU's Approach to AI Governance

The AI Act fundamentally follows a risk-based approach in multiple respects. On one hand, it categorizes various AI applications and imposes different levels of regulatory requirements depending on the category (Floridi et al., 2022; von Welser, 2024, p. 484 ff.).

A key regulatory focus of the AI Act is the formulation of obligations for so-called high-risk systems (Chapter 2 AIA). The operation of these systems is not inherently prohibited, but they must comply with a series of regulatory requirements and function as a product safety regulation (Rohrßen, 2024). Consequently, it is fair to assume that these requirements will have the most impact on the design of future systems available on the market.

The fundamental part of those obligations is the risk management laid down in Art. 9 AIA. Such risk management is regulated in Art. 9 and can be roughly summarized as a continuous, iterative process that identifies, evaluates and, where possible, mitigates risks. For market approval, the risks must be reduced to an "acceptable level". In this context, "known or reasonably foreseeable risks" to "health, safety or fundamental rights" that may arise from the use of the product for its intended purpose or from "foreseeable misuse" must be taken into account (von Welser, 2024).

[47] Regulation (EU) 2024/1689 (Artificial Intelligence Act) further referred to as AIA

In theory, risks to all stakeholders affected by an AI system must be considered. This includes not only the manufacturers, operators, and users of an AI system but also groups who are indirectly impacted by the system without having direct influence on its use. While risks to third parties are therefore theoretically acknowledged, the practical implementation of risk management under the AI Act may fall short in effectively addressing these risks, as it is discussed in the following sections.

5.1.2. Subjectivity in Risk Management

Risk management is not a precise, mathematical process that produces a deterministic outcome. There is always a certain degree of subjectivity on the part of the actor conducting the risk management. This subjectivity influences both the identification of risks – whether they are even considered in the first place – as well as the evaluation of the likelihood of occurrence and the expected damage. This subjectivity particularly affects intangible risks, which are difficult to quantify using discrete categories such as numbers .(Ramnarine, 2015).

The responsibility for conducting risk management falls on the manufacturers, operators, or economic intermediaries laid down in Chapter 3 AIA. This makes sense to a certain extent, as these parties are capable of making concrete changes to an AI system and thus can operationally mitigate risks (Brinker, 2024). However, these actors naturally have vested interests in the design or functionality of the AI system, especially economic interests in bringing an AI system to market or using it in a particular form, which biases the risk management process.

5.1.3. Lack of specificity in risk management requirements

The AI Act is fundamentally designed as a "principle-based" regulation (Schuett, Anderljung, Carlier, Koessler, & Garfinkel, 2024). This high level of abstraction is also evident in the requirements for risk management. No specific methodological guidelines are provided, only eclectic requirements that a risk management process must fulfill.

The lack of specificity becomes critical, however, especially with regard to the types of risks that must be considered. Art. 9 (2) AIA refers to "known or reasonably foreseeable risks" to "safety, health, or fundamental rights." While the obligation to consider fundamental rights is, of course, commendable, there is a danger that this broad category will become a mere compliance checkbox to tick during the risk management process conducted by economic actors. Fundamental rights are universally valid, but due to the high level of abstraction, it is difficult (or nearly impossible) to derive concrete risk scenarios that must be considered for practical risk management. This lack of specificity means that there is a risk that the selection of risks taken into account will remain eclectic. If the manufacturer has no self-interest, there is a danger that third party risks are “forgotten”. Yet even if there are no bad intentions involved, the manufacturers and operators are biased by their own subjective perspective. It’s in the nature of third-party risks, that they are not as obvious for others as they are to the parties directly involved.

Additionally, there may be a lack of methodological expertise on the part of manufacturers or operators of high-risk systems in identifying and evaluating risks to fundamental rights or third parties. While fundamental rights must be universally respected, the methods for identifying or weighing potential infringements are not trivial and are not universally mastered (Janssen, Seng Ah Lee, & Singh, 2022). Given that AI system manufacturers tend to be experts in technical domains, it is likely that they lack the necessary methodological tools or only have rudimentary knowledge of them.

It should be noted that a lack of methodological understanding is no excuse for failing to comply with legal requirements. In case of doubt, an entity is obligated to acquire the necessary methodological knowledge. However, in light of the inherent subjectivity of the risk management process, this is another factor that makes it unlikely that risk management will consistently produce the highest-quality outcomes. Instead, it will likely be conducted at the edge of what is just acceptable.

5.1.4. Principal-Agent Relationship in Risk Management

In essence, it is not the risk owner who decides how a risk affecting him is to be considered, evaluated and, in case of doubt, mitigated, but an actor with a certain vested interest. Since risk management likewise does not lead to an exact result, the actor who carries out the risk management can at least partially influence the result in the direction he desires. Risk management is thus not to be seen as a balancing of interests of all stakeholders involved, but as a means of ensuring minimum standards.

This relationship between legislators and AI manufacturers and operators can be understood through the lens of the principal-agent theory (e.g. Ross, 1973). In this framework, the legislators act as the principals, setting out the requirements and goals (such as safety and protection of fundamental rights, consideration of risks for third parties), while the manufacturers and operators are the agents tasked with implementing these requirements through risk management processes. In principle, this arrangement functions effectively as long as the supervisory authorities, representing the principal, are diligent in their oversight duties (Hussein & Menon, 2003).

However, in practice, challenges arise when the agent's interests diverge from the principal's goals, particularly if the agents are primarily motivated by meeting only the "necessary minimum" requirements. This can lead to a "race to the bottom," where agents do just enough to comply with the law without fully embracing the spirit of the regulation. Such minimal compliance is difficult to counteract once it becomes the norm, even with subsequent legal adjudications or adjustments to the regulatory framework.

5.1.5. Concretization gone wrong

Although the AIA is “principle-oriented”, it contains its own tools for concretizing its abstractly formulated requirements. In addition, a fundamental concretization can develop in practice through the application of law in court rulings, the action of the supervisory authority, or through generally developing conventions (such as public or private standards) (Schuett et al., 2024, p. 33 ff.). While the mechanisms mentioned in the previous sentence are rather indirect in nature, the mechanisms of the AI Act aim at a direct concretization. Accordingly, harmonized standards pursuant to Art. 40 AIA or common specifications pursuant to Art. 41 AIA can be adopted by the Commission by means of an implementing act.

However, neither direct nor indirect concretization takes place in a vacuum, but always against a material technical background. If manufacturers and suppliers have a vested interest in taking third-party risks into account at the smallest justifiable level, this minimal principle will also affect the design of their products. However, by defining the technical facts, they are setting the starting point for the discussion of further concretizations.

This applies in particular to the concretization through case law. The aim of case law is not to identify an optimal risk assessment, but merely to determine inadmissible interpretations. Thus, case law may at most shift the minimum threshold upwards.

However, concretizations through standards (including those officially defined by the instruments of the AI Act) are also influenced by existing technical possibilities, especially if they are already widespread.

To illustrate these effects, the evolution of the infamous "cookie consent" in the context of the GDPR can be used as a somewhat comparable scenario. Even the quite clearly formulated requirements for consent – it must, among other things, be given voluntarily, in an informed manner (i.e. the data subject must know exactly what he or she is consenting to) and unambiguously – led to a series of stylistic bloopers in the implementation of the website operators (who had a corresponding self-interest in ensuring that consent is given) (see e.g. Möller, 2022, p. 455). It took several years for what were actually obviously unlawful consent forms, such as the continued use of the website, pre-selected checkboxes, etc., to be addressed by courts (e.g. EuGH ECLI:EU:C:2019:801). And even eight years after the GDPR came into force, there appears to be little sign of effective enforcement. Even if the most obvious cases have been dealt with in court, the courts are still struggling with more subtle means of manipulation, such as dark patterns (Leiser & Santos, 2023) or simply misleading wording of the consent text. It would be naive to assume that the average internet user actually has an informed idea of what consent to "cookies" actually means.

In order to ensure that third-party risks are adequately taken into account when the AI Act is finalized, similar dynamics must be prevented and, above all, the "minimum standard" set by the manufacturers (Wehkamp, 2022) must not be used as the sole starting point for the discussion.

5.2. Conclusion

5.2.1. Make Third party risks known

It has been shown that third-party risks, for systematic reasons that lie primarily in the risk management carried out by the operators or manufacturers, tend to be given less consideration in risk management.

If this is to be avoided, the AI Act will have to be able to concretize the currently relatively abstract requirements, with the market surveillance authorities playing a central role here. This is especially true for formalized concretization processes through standards or common guidelines. When developing these, care should be taken to ensure that all stakeholders are able to contribute their input, and that civil society is not neglected in favor of the technical community. This is the only way to ensure that manufacturers and providers take due account of third-party risks that do not fall within their own sphere of interest. Although it may not be feasible to get them to do more than "work to rule" and always take the minimalist approach to risk management, it is important that third-party risks are also considered as part of the "work to rule" approach.

In this context, civil society actors have the particular role of making third-party risks generally known. Even manufacturers and operators who are willing to take into account all risks for third parties have a bias due to their subjective perspective and can thus overlook third-party risks. In order to have any chance of being considered, affected stakeholder groups or their representatives must therefore publicly draw attention to any "forgotten" negative effects on themselves or others.

5.2.2. Outlook

As explained, risk management for high-risk AI systems is not about balancing the interests of all stakeholders, but about ensuring minimum standards. This does not necessarily speak against the AI Act as a whole, as minimum standards do not necessarily lead to an optimal result for society as a whole, but are initially a step in the right direction. Nevertheless, the mistake must not be made to see the AI Act as a definitive part of AI regulation due to its generic designation, which takes into account all social impacts.

In order to achieve an appropriate consideration of third-party risks, it is particularly important to concretize the currently still very abstract provisions of the AI Act. In this context, it is important for civil society to draw attention to risks and for the authorities to adequately acknowledge them in the context of concretization.

References

Brinker, N. (2024). Identification and demarcation—A general definition and method to address information technology in European IT security law. *Computer Law & Security Review*, *52*, 105927. https://doi.org/10.1016/j.clsr.2023.105927

Floridi, L., Holweg, M., Taddeo, M., Amaya Silva, J., Mökander, J., & Wen, Y. (2022). capAI - A Procedure for Conducting Conformity Assessment of AI Systems in Line with the EU Artificial Intelligence Act. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4064091

Hussein, K., & Menon, A. (2003). The principal-agent approach and the study of the European Union: Promise unfulfilled? *Journal of European Public Policy*, *10*(1), 121–139. https://doi.org/10.1080/1350176032000046976

Janssen, H., Seng Ah Lee, M., & Singh, J. (2022). Practical fundamental rights impact assessments. *International Journal of Law and Information Technology*, *30*(2), 200–232. https://doi.org/10.1093/ijlit/eaac018

Leiser, M., & Santos, C. (2023). Dark Patterns, Enforcement, and the emerging Digital Design Acquis: Manipulation beneath the Interface. *BILETA Special Issue*, *1*(15). Retrieved from https://ssrn.com/abstract=4431048

Möller, C. C. (2022). Dark Patterns in Consent-Bannern. *Verbraucher Und Recht*, *37*(12), 449–458.

Ramnarine, E. (2015). Understanding Problems of Subjectivity and Uncertainty in Quality Risk Management. *Journal of Validation Technology*, *21*(4).

Rohrßen, B. (2024). KI & CE – Die KI-VO, das Produktsicherheitsrecht für Künstliche Intelligenz. *Zeitschrift Für Produkt Compliance*, *1*(3), 111–123.

Ross, S. A. (1973). The economic theory of agency: The principal's problem. *The American Economic Review*, *63*(2), 134–139.

Schuett, J., Anderljung, M., Carlier, A., Koessler, L., & Garfinkel, B. (2024). *From Principles to Rules: A Regulatory Approach for Frontier AI* (Version 1). Version 1. arXiv. https://doi.org/10.48550/ARXIV.2407.07300

von Welser, M. (2024). Die KI-Verordnung – ein Überblick über das weltweit erste Regelwerk für künstliche Intelligenz. *Gewerblicher Rechtschutz Und Urheberrecht in Der Praxis*, *16*(15), 485–488.

Wehkamp, N. (2022). Internalization of Privacy Externalities through Negotiation: Social costs of third-party web-analytic tools and the limits of the legal data protection framework. *Companion Proceedings of the Web Conference 2022*, 525–533. Virtual Event, Lyon France: ACM. https://doi.org/10.1145/3487553.3524631