

Jurgen Goossens
Esther Keymolen
Antonia Stanojević *Editors*

Public Governance and Emerging Technologies

Values, Trust, and Regulatory
Compliance

OPEN ACCESS

 Springer

Public Governance and Emerging Technologies

Jurgen Goossens • Esther Keymolen •
Antonia Stanojević
Editors

Public Governance and Emerging Technologies

Values, Trust, and Regulatory Compliance

Editors

Jurgen Goossens
School of Law
Utrecht University
Utrecht, The Netherlands

Esther Keymolen
Tilburg Law School
Tilburg University
Tilburg, The Netherlands

Antonia Stanojević
Tilburg Law School
Tilburg University
Tilburg, The Netherlands



ISBN 978-3-031-84747-9

ISBN 978-3-031-84748-6 (eBook)

<https://doi.org/10.1007/978-3-031-84748-6>

This book was financially supported by the Dutch Research Council (NWO) as part of the project “The Role and Responsibilities of Public Actors in Distributed Networks. Transparency, Trust and Legitimacy by Design” [project no. MVI.19.040] within the framework of the program “Responsible Innovation. Designing for public values in a digital world.”

© The Editor(s) (if applicable) and The Author(s) 2025. This book is an open access publication.

Open Access This book is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this book are included in the book's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the book's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

If disposing of this product, please recycle the paper.

Acknowledgments

This book was financially supported by the Dutch Research Council (NWO) as part of the project “The Role and Responsibilities of Public Actors in Distributed Networks. Transparency, Trust and Legitimacy by Design” [project no. MVI.19.040] within the framework of the program “Responsible Innovation. Designing for public values in a digital world.”

We are grateful to Charlotte van Oirsouw, Juliette Ermers, Marc Hanna, Dennis Wegink, and Circe Janse for their support with conceptualizing the edited volume and preparing the manuscript.

We are also grateful to all contributors for their sustained efforts throughout the writing process and their engagement in the “Public governance and emerging technologies: values, trust and compliance by design” workshop that served as a preparation for the publication of this edited volume.

Contents

Governance by Technology, Governance of Technology: The Quest for Public Values, Trust, and Regulatory Compliance	1
Antonia Stanojević, Jurgens Goossens, and Esther Keymolen	
Part I Governance: By and of Technology	
Municipalities Enacting Algorithms: A Typology of Dutch Municipal Strategies for Leveraging Algorithmic Systems	19
Maranke Wieringa	
The Role of Technology in Citizens’ Right to Good Administration: Examining the Impact of Smart Governments	43
Barbara Lazarotto	
Explainable AI Is No Silver Bullet: Towards a Contextual Understanding of Appropriate Reliance on AI in Law Enforcement	61
Merijn Bruijnes, Stephan Grimmelijsen, and Marcel Robeer	
European Digital Infrastructure Consortium (EDIC): A New Governance Framework for the European Blockchain Services Infrastructure (EBSI)	83
Evrin Tan and Daniël Du Seuil	
Privacy-By-Design in the Belgian Public Sector	103
Kristof Verslype and Bart De Decker	
Our Stakes in Data: How Do We (Re)Gain Democratic Control Over Digital Practices?	131
Barbara Prainsack	
Part II Regulatory Compliance: “Here We Must Run as Fast as We Can, Just to Stay in Place”	
Regulatory Equivalence in Blockchain Systems: The Role of Public Values and Legitimacy	151
Primavera De Filippi and Morshed Mannan	

The Need for Responsible Use of AI by Public Administration: Algorithmic Impact Assessments (AIAs) as Instruments for Accountability and Social Control	179
Julia Iunes Monteiro	
As Much as Faith: A Speculation on Quantum Computing with Fiduciary Law in Public Governance	217
Robert Herian	
The Rule of Tech? Submitting Legislation to Digital Constraints	239
Sabrina Kutscher and Johan Wolswinkel	
The Shaping of AI Policies in Latin America: A Study of International Influence and Local Realities	263
María Lorena Flórez Rojas	
Part III Public Trust and Values: In Technology We Trust?	
Empathy as a Public Value: Overcoming Administrative Vulnerability and Rehumanizing (Digital) Government	291
Malou Beck and Sofia Ranchordás	
Civic Vice in Digital Governance	315
Wessel Reijers	
Algorithmic Governance and Social Vulnerability: A Value Analysis of Equality and Trust	335
Antonia Stanojević	
The Frameworks of Trust and Trustlessness Around Algorithmic Control Technologies: A Lost Sense of Community	355
Balázs Bodó and Linda Weigl	
Learning-Loop Bureaucracies: A Configuration Approach to AI, Law, and Organizations	381
Stavros Zouridis	

Governance by Technology, Governance of Technology: The Quest for Public Values, Trust, and Regulatory Compliance



Antonia Stanojević, Jurgen Goossens, and Esther Keymolen

Abstract This chapter situates a collection of interdisciplinary perspectives gathered in the edited volume ‘Public Governance and Emerging Technologies: Values, Trust, and Regulatory Compliance’ within the growing literature on public governance and emerging digital technologies. By using topical examples and a sociomaterial approach, it reflects on the observed challenges in the governance of and governance by emerging technologies. Following, the chapter approaches these issues from a regulatory perspective, discussing regulatory compliance and identifying opportunities for future legislative developments. Thereafter, a normative analysis centers public values and trust, drawing on lessons from public governance theory and practice. Overall, the chapter emphasizes the rapid nature of developments in governance by and of technology, and far-reaching social, political and material consequences of these developments. By identifying opportunities and challenges related to emerging digital technologies, it serves to inform practitioners in the field, as well as to identify opportunities for further research. Throughout the chapter, the central role of public values and trust is emphasized as normative ideals that should lead the developments in legislation and public governance.

In the complex network society, public actors are increasingly experimenting with utilizing emerging digital technologies, such as artificial intelligence and blockchain. In conjunction with this process, public authority is no longer solely exercised in a unilateral manner but is increasingly also carried out through governance networks of public and private actors. ‘Disruptive’ or ‘emerging’ digital technologies, are often portrayed as highly promissory, enshrined with the hope to transform established societal, economic, or governmental practices. These promises range from making processes more efficient and transparent to developing privacy-friendly ‘data solutions.’ At the same time, the use of networked and data-driven technologies to mediate the interaction between public actors and citizens

A. Stanojević (✉) · E. Keymolen

Tilburg Law School, Tilburg University, Tilburg, The Netherlands

e-mail: a.stanojevic@tilburguniversity.edu; e.l.o.keymolen@tilburguniversity.edu

J. Goossens

School of Law, Utrecht University, Utrecht, The Netherlands

e-mail: j.goossens@uu.nl

© The Author(s) 2025

J. Goossens et al. (eds.), *Public Governance and Emerging Technologies*,
https://doi.org/10.1007/978-3-031-84748-6_1

challenges and transforms public values such as transparency, accountability, fairness, and trust. Public values of the rule of law do not necessarily align with values in network governance such as efficiency and scalability, resulting in value conflicts. Due to value conflicts and unclear legal norms citizens might experience vulnerability and uncertainty, negatively impacting the trust relationship with their government. This is particularly problematic in the context of declining political trust over the past decades, as observed in some European countries.¹ Moreover, the increased use of complex network governance and opaque emerging technologies to exercise public authority may cause uncertainty for citizens in terms of legal protection, particularly as legislative developments struggle to keep up with the pace of technological innovation.

This book brings forth a multilevel, interdisciplinary analysis of the described trends at the intersection of public governance and emerging digital technologies. Based largely on qualitative and conceptual research, this edited volume discusses challenges and opportunities that appear in the space where public governance meets emerging digital technologies. Using topical examples, it offers a balanced collection of critical and optimistic considerations, thereby juxtaposing challenges and opportunities. To this end, the volume gathers perspectives from public administration, sociology, political science, computer science, law, philosophy, and science and technology studies.

Part I ‘Governance – By and of Technology’ of this book shows how automatization by algorithmic systems, the networked nature of distributed technologies such as blockchain, and the data-driven use of AI in public governance proliferate hyper-connectivity and hyper-complexity. Part II ‘Regulatory compliance - “Here we must run as fast as we can, just to stay in place”’ turns to the analysis of the aptness of the existing legal framework to developments in the field of emerging digital technologies and discusses legislative adaptations needed to adequately safeguard public values by regulating the use of such technologies in public governance. Expanding on this, Part III ‘Public trust and values – in technology we trust?’ highlights public values as the common thread underpinning the book, and the key considerations that should guide the application of emerging digital technologies in public governance. As such, the edited volume starts by describing examples of public governance of and by technology, proceeding to outline regulatory considerations of such increasingly prevalent practices, finally concluding with value analyses and opening space for further normative discussion and the creation of new social imaginaries. The following sections of the introduction aim to briefly introduce each chapter and reveal the underlying connections that weave these chapters into an overarching narrative.

¹ See van der Meer and van Erkel (2024); Cole et al. (2022); Torcal and Christmann (2021).

1 Governance: By and of Technology

Algorithmic systems are increasingly used to innovate administrative decision-making procedures, as shown in the typology of municipalities' adoption of algorithmic systems developed by Wieringa in the first chapter. In doing so, this chapter takes a sociomaterial perspective that includes material, social, cultural, and discursive aspects, considering them inseparably interwoven. Based on a four-year long qualitative study of the implementation of algorithms into public administration procedures by Dutch municipalities, this contribution introduces a typology of different municipal enactment strategies. The findings indicate that municipalities can be divided along two axes: first, by the rate of adoption of algorithmic systems (slow and fast); and second, by the degree to which they dedicate resources to responsible and accountable algorithmization (more or less conscientious). These two dimensions define four archetypes: frontrunners, fast fashionistas, observers, and slowpokes. The four archetypes of introducing algorithms in public administration have different goals, advantages, and disadvantages, and may be appropriate in different sociomaterial contexts. However, dedication to a conscientious and reflexive approach to algorithmization is always recommendable. Therefore, the observer strategy, which minimizes risks and draws on best practices, may be a good option for most municipalities. The other conscientious strategy, namely frontrunner, may not be appropriate for most municipalities, as it requires substantial effort and resources. On the other hand, strategies low on conscientious reflexivity are not recommendable in the current political climate around algorithmization. However, applying a high rate to low conscientiousness (the fast fashionista strategy) is likely even more problematic than lower conscientiousness with a lower rate of adoption (the slowpoke's strategy). The chapter opens the space for further research on differences between strategies of algorithm implementation in public administration, highlighting the need for context-specific recommendations regarding best practices.

In the second chapter, Lazarotto reflects on "smart governments" as a potential example of algorithmization practices that could benefit from more conscientious and slower implementation. The phenomenon of "smart governments" is defined as the general tendency towards automatization of public services. While the narrative that motivates the widespread implementation of emerging technologies in public governance places an emphasis on efficiency and optimization, focusing thereby on the governments' perspective, this chapter emphasizes the citizens' perspective. As such, this chapter aims to scrutinize the proliferation of "smart governments" in the European Union, analyzing their compatibility with the right to good administration as enshrined in the EU Charter of Fundamental Rights. As a result of this analysis, the chapter proposes a multi-layered framework focused on citizen empowerment. It emphasizes essential rights such as transparency and the right to be heard. In addition to fundamental rights, it reflects on the implications of the recently adopted AI Act for "smart governments". The chapter demonstrates the need for a set of governance structures and safeguards in step with the rapid algorithmization of public governance.

Grimmelikhuijsen, Robeer and Bruijnes deepen the discussion on concerns over the societal consequences of increased use of algorithms in public governance in the third chapter. Focusing on street-level decision making in public administration, this chapter presents a contextual model of trustworthiness determinants of AI. Specifically, it studies the choices police officers make when using AI recommendations in their daily tasks, and particularly the role of explainable AI (XAI) in those choices. The chapter relies on empirical data collected through five studies, two qualitative and three quantitative, conducted at the Netherlands Police Lab for Artificial Intelligence. Theoretically rooted in organizational science and sociology, the chapter explores appropriate decision-making in the context of AI recommendations used by law enforcement. In general, a prominent concern around AI-supported decision-making is the tendency of humans towards overreliance on AI recommendations, as indicated by findings of experimental studies.² To counter this issue, this chapter highlights the need for appropriate reliance, defined as an appropriate response to AI recommendations as neither stubborn disagreeableness, nor complacency in agreeing with AI recommendations. The chapter argues that XAI leads to appropriate responses by providing context for the recommendation, based on which the street-level bureaucrat decides whether it should be accepted or rejected, thereby relying on additional information rather than (mis)trust. However, the findings show that XAI is not a guarantee of appropriate reliance, but rather one of the factors that can support appropriate reliance. Based on this, a contextual model is proposed to help identify individual and task characteristics that foster appropriate reliance on XAI recommendations. The chapter highlights the need for future research to focus on organizational and occupational context, to deepen the understanding of appropriate reliance and further inform the contextual model.

While the first three chapters focused on the use of emerging digital technologies in public governance (governance *by* technology), the fourth chapter turns to a discussion of governing the necessary infrastructure (governance *of* technology). In this chapter, Tan and Du Seuil use the example of European Blockchain Services Infrastructure (EBSI) to analyse the processes that occur when proliferation of emerging technologies demands from public actors the development of digital infrastructure. These emerging digital infrastructures typically require new forms of governance, often relying on intricate cooperations between public and private actors. In these cases, public actors must often rely on the expertise of the private sector, leading to new forms of public-private partnerships and network governance, the complexity of which may give rise to concerns around transparency and accountability. This chapter uses qualitative data (desk research and expert interviews) to explore the inception of the European Digital Infrastructure Consortium (EDIC, formally established in 2024 under the name Europeum), a novel governance model of the European Blockchain Services Infrastructure (EBSI). The chapter draws on the insights from management science, namely institutional

²See Klingbeil et al. (2024).

isomorphism theory³ to explain organizational choices driving the formation of EDICs. It concludes that, while Europeum emerged as a traditional form of governance due to the inertia of normative and mimetic isomorphism, its further development may turn to more decentralized governance models aligned with the original philosophical underpinnings of the blockchain technology, such as Decentralized Autonomous Organizations (DAOs).

Following, Chap. 5 examines the governance of data in the public sector. In this chapter, Verslype and De Decker introduce the use case of joining dispersed data to gain additional insight, by discussing two specific healthcare-related cases from Belgium. Protection of individual privacy is of key importance when joining data and becomes particularly apparent when applied to sensitive data, such as data reflecting biometric and health related personal information. It is precisely for privacy reasons that data joining is not commonly done at scale, but rather on project basis, where each research question drives a new process of joining only the necessary personal data, in line with GDPR principles of proportionality and data minimization. However, at times research questions require joining large datasets to discover unknown patterns and predict or prevent outcomes based on such patterns. Such are the cases presented in Chap. 5, where joining is conducted towards uncovering patterns in the epidemics of diabetes and cancer risk of multiple sclerosis patients. Using these two cases as examples, this chapter discusses the advantages and challenges of two cryptographic innovations based on joining and pseudonymizing data. The chapter concludes with a reflection on the juxtaposed public values of privacy and transparency and argues that emerging cryptographic methods can help safeguarding both simultaneously.

Finally, Chap. 6 acts as a bridge between Part I (governance) and Part II (regulatory compliance). In this chapter, Prainsack acknowledges that many of the societal harms associated with digital practices are not at odds with the law.⁴ Two explanations are offered for the observed regulatory compliance of negative effects of digitalization, namely lobbying and the dominance of fair market competition as the regulatory framework. Therefore, Chap. 6 emphasizes the need to move beyond the analysis of existing regulatory compliance to (re-)establish democratic control over digital practices and infrastructure as a common resource. The chapter proposes two such paths towards democratic control: firstly, it advocates for strengthening existing and adopting new rights; and secondly, it promotes democratization of control over data and digital infrastructures. Prainsack argues that these two goals cannot be achieved simply through adjustments in existing regulatory instruments, pointing out failures of previous legislative tweaks to effectively address the core societal issues. Instead, achieving both necessitates a new conceptualization of democratic data governance. In this vein, the chapter outlines a three-layer model of control over data, consisting of individual-, community- and national-level control. Additionally, Prainsack asserts that, in parallel to democratic control over data,

³See DiMaggio and Powell (1983).

⁴See also Prainsack (2015).

democratic control should be established over digital infrastructures (cf. Chap. 4), whereby digital infrastructures should be treated as public infrastructures similar to roads or train tracks. Part I of the book thus sets the stage by identifying the key issues in governance by technology (Chaps. 1–3) and governance of technology (Chaps. 4–6).

2 Regulatory Compliance: “Here we Must Run as Fast as we Can, Just to Stay in Place”

These evolutions and the associated concerns draw societal, political, and scholarly attention to regulatory compliance of the current and potential uses of emerging technologies, as demonstrated in Part II ‘Regulatory compliance—“Here we must run as fast as we can, just to stay in place”’ of the book.

In this vein, De Filippi and Mannan (Chap. 7) discuss regulatory compliance of blockchain technology through the lens of regulatory equivalence “to stay in place”. First, the concepts of ‘functional equivalence’ and ‘regulatory equivalence’ are distinguished from the notions of technological neutrality and regulation by design. Functional equivalence pursues regulatory compliance by establishing equivalence between the objective of existing legal formalities and the technological guarantees provided by a new technological artifact. Regulatory equivalence aims for alignment with the overarching policy objectives embedded within specific legal provisions rather than strict adherence to established legal formalities. According to De Filippi and Mannan, regulatory equivalence represents an evolution of functional equivalence towards a prefiguration of a future regulatory environment in which public policies and technological innovation harmoniously coexist.⁵ Second, they identify opportunities and challenges to achieve regulatory equivalence, using examples of blockchain-based notarization processes, DAOs, and privacy pools. Formal recognition of regulatory equivalence by lawmakers and public authorities remains particularly challenging. Third, the authors illustrate the pivotal role of public values in the quest for legitimacy. Based on different conceptions of legitimacy as well as the interconnectedness between endogenous and exogenous legitimacy, the authors show how the political system of public authorities and the Web3 ecosystem have both convergent and divergent values. Both systems strive for input and procedural legitimacy. Yet, substantive values are more likely to diverge resulting in value clashes and diverse ways to pursue legitimacy. Such a clash of values, however, may incite inappropriate forms of regulation and hinder regulatory equivalence from being officially recognized. On a more positive note, the interplay between the political system and emerging technological ecosystems could bring implicit values to the surface that are not explicitly embedded in the applicable legal

⁵ See also Collomb et al. (2019).

framework. Such an interaction could help practitioners to design technological systems in ways that better provide regulatory equivalence.

As part of the quest to guarantee the responsible use of AI in public governance, Iunes (Chap. 8) assesses the impact of algorithmic governance on human rights building on examples from the Global North and Global South alike. The increased use of AI in decision-making by administrative authorities has a substantial impact on citizen's fundamental rights. Iunes argues that the indiscriminate use of AI undermines the legality, transparency, and justification of administrative acts. Notwithstanding the early stages of regulation and practice, AIAs are increasingly proposed as a key accountability mechanism.⁶ Both the AI Act in the EU and the Brazilian Legal Framework for Artificial Intelligence determine that AIAs are mandatory for high-risk systems that pose an elevated degree of harm to people's health, safety, or fundamental rights. Applying AIA before the introduction of an algorithmic system in the real world can prevent risks and provide an evaluation of its societal implications. Demonstrating the significant role AIAs can play to preserve trust in government, the rule of law, and fundamental rights, Iunes explores the fundamentals, the challenges, stages, and requirements of AIA to become an effective mechanism for algorithmic accountability by the public administration. She argues that the establishment of binding rules, a supervisory authority, its scope of application, as well as the regulation of its implementation process are important criteria for the institutionalization of AIA. According to Iunes, AIA should not only be used as a risk management mechanism but must also be considered as an instrument of reflection and public debate on the adequacy of an AI system in a concrete social context. It must serve as a source of information enabling public scrutiny of algorithmic policies implemented by the public administration. To enable effective democratic and social control, Iulia thus advocates AIAs must include participatory processes and mechanisms for public consultation and contestation. However, participation is rarely observed in practice, neither is it a requirement in, for instance, the recent AI Act.

Thereafter, Herian (Chap. 9) explores how fiduciary law may offer regulatory tools for safeguarding core public values in the face of quantum computing and its emergent use in governance. By harnessing the laws of quantum mechanics, quantum computing can address problems too complex for classical computers and do so with much greater speed and efficiency. Nonetheless, Herian warns that quantum computing presents a level of technological innovation threatening to outpace human know-how, understanding and subsequently effective control. As an increasingly heavy techno-social burden rests on citizens and knowledge gaps expand, there is a need to have belief or faith—what Herian refers to as quantum fide (as much as faith)—in the stakeholders of technological systems. Hence, quantum computing extends an invitation to embrace “known unknowns” and leaps of faith. Safeguarding core public values and compliance with legal norms are at stake when designing and implementing quantum computing as part of a broader socio-technical

⁶ See Ada Lovelace Institute et al. (2021).

infrastructure used in public governance. Although experts play an important role, he argues that social democracy and governments must determine policy, define rules, ensure oversight and accountability of stakeholders, thereby relieving the techno-social burden on citizens. Therefore, this chapter explores the concept of fiduciary law as a body of theory and practice directed towards acting on behalf of others and helping relieve the burden of expertise and know-how of increasingly complex emerging technologies.⁷ Herian argues that fiduciary obligations, duties, and responsibilities as part of meaningful and enforceable techno-social governance frameworks could show the way how emerging technologies in public governance might proceed from a promissory idea to development and deployment in a legally, ethically, and socially acceptable way.

Next, Kutscher and Wolswinkel (Chap. 10) demonstrate that not only ‘by design’ approaches of regulatory compliance are gaining ground but there is also increased attention for adopting legislation which meets digital demands. They discuss ways to develop ‘digital-ready’ legislation that can adequately address the fast-paced evolutions and co-existence of technology, societal challenges, and legal remedies. The exploration of machine-consumable legislation in three frontrunner jurisdictions (Denmark, Aotearoa (New Zealand), and the Netherlands) shows the potential impact on the legislative process. These jurisdictions are experimenting with digital-ready legislation that facilitates the implementation of legislation through digital systems, thereby transforming the legislative process itself. After comparing several forms of digital-ready legislation, Kutscher and Wolswinkel assess their consequences for the technocracy-democracy divide as well as the virtues traditionally associated with the legislative process. On the one hand, digital-ready legislation may strengthen the virtues associated with the form or outcome of legislation, particularly the purposive generality and transparency of the legislative output. On the other hand, digital-ready legislation challenges virtues associated with the process of enacting legislation, such as democracy, deliberation, and transparency. Even though these virtues thus still have an important role to play, the authors illustrate that legislation is increasingly not merely subjected to the ‘rule of law’ but also to the ‘rule of tech’. They advise the formal legislator to make digital constraints imperative in the legislative process to avoid technocracy to prevail over democracy but encourage their fruitful co-existence.⁸ Ensuring digital compliance of legislation ‘by design’ necessitates, among others, collaboration in multidisciplinary teams of lawyers, policymakers, and ICT professionals when drafting legislative proposals.

Flórez Rojas (Chap. 11) concludes the regulatory compliance section of the book by reflecting on lessons learned from approaches in the Global South and exploring possibilities beyond the current approaches. She identifies the widespread mention of ethical guidelines and analyses the main regulatory paths. Several Latin American have taken as regards AI governance. Based on a qualitative methodology, she assesses AI policies and strategies in Argentina, Colombia, Chile, Costa Rica, and

⁷ See Miller and Gold (2015).

⁸ See Gilley (2017), p. 18.

Brazil, with a focus on the regulatory influence of international organizations in the development of these national policies. Thereafter, the author delves into the limitations and challenges concerning the primary influence of external parties in national strategies in view of the region's diverse cultural, political, and socioeconomic landscape. The region has acknowledged the need for a tailored regulatory framework that goes beyond legal rules for AI, broadening its focus to encompass initiatives promoting digital literacy, technological infrastructure, research, public participation, and diverse stakeholder involvement. Nonetheless, Flórez Rojas observes that the analyzed countries gave little space to assess social and cultural challenges within their national AI strategies, whereby the influence of international values diminished their own national values and cultural approaches.

Due to the underrepresentation of Latin American perspectives in global AI and ethics debates, Flórez Rojas raises concerns about the adequacy of existing frameworks to meet regional needs. The development of AI national strategies is at varying stages across Latin American countries, with some responding to market pressures, potentially prioritizing economic considerations over ethical ones. Therefore, Flórez Rojas advocates the implementation of networked governance, going beyond 'multistakeholderism', across all aspects of AI governance mechanisms, emphasizing the importance of trust, legitimacy, learning, power symmetry, and fairness within these frameworks.⁹ According to the author, such an approach must recognize the interrelation of external and internal influences in shaping governance outcomes, emphasizing the need for collaborative efforts among international organizations, governments, and various stakeholders.

In conclusion, Part II turns the focus towards regulatory solutions to issues identified in Part I. Part II explores the compatibility of technology with existing regulations (Chaps. 7 and 8), existing legal tools that could be innovatively applied for successful regulation of emerging technologies (Chap. 9), and approaches to updating existing or creating new legislation for regulation of emerging technologies (Chaps. 10 and 11).

3 Public Trust and Values: In Technology We Trust?

Examples given throughout Parts I and II of the book show that emerging technologies represent a challenge for important values, particularly when used in public governance. While socio-ethical considerations on upholding public values in a digital world are at the heart of all book chapters, Part III 'Public trust and values – in technology we trust?' puts these public values and trust center stage. Following Part II, which brings forth a practical discussion of regulatory solutions, Part III advances a conceptual and normative discussion, focusing on trust and other fundamental public values that are strained by technology-induced solutions. Several

⁹ See Wegner and Verschoore (2022).

authors specifically address the Dutch childcare benefits scandal as a clear, yet painful, example of the negative societal impact of algorithmic decision-making. To effectively safeguard these values and prevent similar scandals, it is essential to rethink and reconceptualize what these values truly mean within the context of network governance.

In Chap. 12, Beck and Ranchordás advocate for the recognition of empathy as a key public value in the context of algorithmic decision-making. Empathy, understood as the ability to relate to others, serves as a crucial antidote to the administrative vulnerabilities that groups of citizens encounter in the context of a networked and data-driven government. Building on vulnerability theory,¹⁰ Beck and Ranchordás identify several vulnerability traits linked to digital government, such as its situatedness, multi-dimensional nature, and asymmetric power relations. Fostering empathy should make the government more responsive to the often complex, vulnerable situation of citizens. It should motivate the government to cater to the specific needs of these citizens. Holding up empathy as a guiding value for policymaking and allowing for specific assessments when needed, the authors present an argument in favor of rehumanizing public administration, improving citizen-government interactions, and restoring and strengthening trust in government.

Whereas Beck's and Ranchordás' starting point is public values, which are inherently focused on the organizational and institutional level, Reijers (Chap. 13) begins his analysis at the micro-level, through the lens of civic virtue. Civic virtues (and vices) are dispositions that citizens, and particularly civil servants, cultivate to support (or undermine) the proper functioning of government. Incidents involving algorithmic decision-making often highlight civil servants or technology as the primary culprits, attributing blame to either ill-intentioned individuals or flawed systems. Reijers, however, argues that virtues and vices are always technologically and institutionally mediated. Moral dispositions matter, but we should not ignore the pre-sorting structures in which they emerge. Reijers identifies three risks this interplay of moral dispositions and technological and institutional systems brings forth: (1) increasing domination, leading to servility, (2) increasing conformity, leading to presumptuousness, and (3) the transformation of political action into non-political action, leading to political recalcitrance. To address these risks, it is not sufficient to merely focus on the technological side of things, which is often captured by the catch-all concept of "by design" strategies. It involves a broader understanding of principles like "explainability" as they apply to governance. This principle must encompass both institutions and the technologies they use, ensuring they foster, rather than undermine, practical wisdom.

Through the conceptual lens of experimentation, Stanojević draws our attention to the recurring issue of targeting socially vulnerable individuals and communities when introducing new algorithmic governance tools in Chap. 14. Social vulnerability entails lower resilience to adversity and is closely related to lower socio-economic status. Examples of social vulnerability include, but are not limited to

¹⁰ See Fineman (2008) and Fineman (2021).

senior, disabled, homeless and undocumented individuals, ethnic and racial minorities, and members of rural and/or remote communities. These already vulnerable individuals and groups are exposed to algorithmic governance experimentation which, due to novelty and outcome uncertainty, holds the potential for creating adversity, making them even more susceptible to harm. The chapter connects the existing literature on socio-technical experimentation¹¹ and algorithmic automation of inequality,¹² expanding on these findings by drawing attention on the detrimental effect such trends can have on public values. Furthermore, this chapter acknowledges that members of vulnerable communities typically hold lower levels of social and political trust, while also seemingly enjoying lower levels of trust from their governments. Stanojević concludes with a policy recommendation that strongly advises against the experimentation with algorithmic governance on vulnerable populations because this deteriorates equality and trust – public values already hindered among the vulnerable members of society. Therefore, the chapter concludes that social vulnerability should be considered a counterindication for the nascent use of algorithmic governance, particularly in the shape it tends to take currently—that of surveillance and control.

In Chap. 15, Bodó and Weigl reflect on the goal of trustless systems, such as blockchain, aiming to replace traditional social and institutional mechanisms of producing trust. These trustless systems are introduced to automate tasks by minimizing human intervention and reducing reliance on the goodwill or competence of intermediaries, institutional actors, or community members. However, by exploring decentralized identity management as a key example of a zero-trust use case, the authors compellingly argue that trust cannot be entirely replaced by a technocentric approach. While technological interventions may indeed enhance efficiency and transparency, they do not render trust obsolete. In fact, intermediaries and public institutions are essential for effective adoption of such systems. Addressing issues of perceived or actual untrustworthiness within these institutions requires deeper engagement in the production of trust, rather than abandoning them for a purely protocol-driven, self-sovereign alternative.

In the last chapter, Zouridis reflects on the evolution of the interplay between technology, legislation, and large-scale bureaucracy.¹³ He outlines two successive configurations observed in large-scale bureaucracies over the past five to seven decades. Beginning with a street-level-centered configuration, Zouridis traces the shift to a production-centered configuration, which has evolved into an information-centered one. These shifts demonstrate how the design and substance of legislation have co-evolved with the organizational principles and technologies used by large-scale bureaucracies. Although AI has only recently been integrated into governmental operations, making it challenging to fully describe the future of its use in bureaucracies, Zouridis offers an initial glimpse into the potential changes AI may

¹¹ See van de Poel (2017).

¹² See Eubanks (2018).

¹³ See also Bovens and Zouridis (2002).

trigger in the configuration of legislation and organizational structure and a brief reflection on the constitutional implications of these emerging transformations. Part III, and the book itself, are thus concluded by Prof. Zouridis' account of lessons learned from public governance theory and practice.

4 Looking Back, Moving Forward

This edited volume identifies the conundrums that emerging technologies pose on public governance (Part I). It acknowledges that challenges arise not only when emerging technologies are implemented in public governance (governance by technology), but also when it comes to decisions that governments and international institutions must make regarding governance of technologies, for example technological infrastructures (governance of technology). Thereafter, identified issues are analyzed through the lens of regulatory compliance (Part II) in search of potential regulatory solutions. Finally, Part III expands this search from the legislative to the conceptual realm and seeks for solutions to identified problems in prioritizing public values in public governance. The book shows that the success of using emerging technologies in public governance is largely dependent on available expertise and choices made by key stakeholders in public administration, legislative and regulatory bodies, and tech companies. It also becomes clear that it is crucial to uphold public values and legal rules in a way that sufficiently considers the citizens' perspective when using and regulating emerging technologies in public governance. The chapters explore and offer guidance on how emerging technologies in public governance might proceed from a promissory idea to development and deployment in a legally, ethically, and socially acceptable way.

Despite digital technologies often being perceived as immaterial, due to the opaque nature of their infrastructure and intangible computing processes, it is crucial to consider the material aspects of digital technologies, e.g. how they depend on raw materials,¹⁴ energy consumption¹⁵ and spatial embeddedness.¹⁶ The material aspect of emerging digital technologies is recognized in the edited volume, which takes a sociomaterial approach to analyzing the intersection between emerging technologies and public governance. The sociomaterial approach emphasizes the assemblages of the social and material factors, which interact and in doing so bring about evolving patterns of actions.¹⁷ Such sociomaterial approach is particularly evident in Part I, which demonstrates the interconnectedness of the social and material forces in shaping human behavior (Chaps. 2 and 3), as well as the formation and functioning of institutions (Chaps. 1, 4, 5 and 6). In doing so, Part I sheds light on

¹⁴ See Taffel (2012).

¹⁵ See Berkhout and Hertin (2004).

¹⁶ See Atkins (2021).

¹⁷ See Yeow and Faraj (2014).

how emerging technologies as material objects and technological infrastructures interact with socio-political processes, as well as the roles and responsibilities of public actors in these interactions. Building on this, Part II considers legislative adaptations to the new sociomaterial landscape developing through interactions between emerging technologies and society. While Chaps. 7 and 8 of Part II look to the past and present in understanding regulatory compliance of emerging technologies, Chaps. 9, 10 and 11 turn to the future, imagining ways in which legislation could adapt to emerging sociomaterial assemblages. Similarly, Part III contrasts the past and present public governance by and of emerging technologies with normative ideals, progressing as such in each chapter from looking back to moving forward and drawing lessons that can help uphold public values in public governance theory and practice.

Throughout the edited volume, examples from public governance illustrate the grand challenges facing modern society. These are so-called wicked problems: issues that are unique and at the same time inextricably connected to other issues; with stakes so high that a trial-and-error approach is deemed inappropriate, and yet they are void of “correct” solutions; making them as such difficult to define and even more difficult to solve.¹⁸ Many current public governance issues thus fall under the “wicked” category, e.g. terrorism and crime,¹⁹ urbanization and inequality,²⁰ or climate change, also known as a “super wicked problem”.²¹ Against this backdrop, emerging digital technologies are often, due to their immense potential for sociomaterial change, portrayed as potential solutions to the seemingly ever-more complex societal issues. However, as the outlined problems remain wicked and new technology-related, sociomaterial problems emerge, public governance scholars and practitioners are faced with an important yet challenging question: to what extent are emerging technologies a solution, and to what extent a problem in its own right?²² This question is tackled in the book through a critical, yet nuanced analysis that acknowledges the high risk—high reward potential of emerging digital technologies application in public governance. In doing so, chapters provide concrete policy recommendations and legal guidance on governance by and of emerging technologies that prioritize public values, trust and regulatory compliance.

References

Ada Lovelace Institute, AI Now Institute, & Open Government Partnership (2021) Algorithmic Accountability for the Public Sector. Open Government Partnership. <https://www.opengov-partnership.org/documents/algorithmic-accountability-public-sector/>

¹⁸ See Rittel and Webber (1973).

¹⁹ See Kim and Zhang (2016).

²⁰ See Zhang and Kim (2016).

²¹ See Levin et al. (2012).

²² See Zhang and Kim (2016); Winkel (2021).

- Atkins E (2021) Tracing the 'cloud': Emergent political geographies of global data centres. *Polit Geogr* 86:102306. <https://doi.org/10.1016/j.polgeo.2020.102306>
- Berkhout F, Hertin J (2004) De-materialising and re-materialising: digital technologies and the environment. *Futures* 36(8):903–920. <https://doi.org/10.1016/j.futures.2004.01.003>
- Bovens M, Zouridis S (2002) From street-level to system-level bureaucracies: How information and communication technology is transforming administrative discretion and constitutional control. *Public Adm Rev* 62(2):174–184. <https://doi.org/10.1111/0033-3352.00168>
- Cole A, Stafford I, Heinz D (2022) Democratic decline? Civil society and trust in government. In: *Civil society in an age of uncertainty*. Policy Press, pp 133–162. <https://doi.org/10.1016/j.cities.2013.03.010>
- Collomb A, De Filippi P, Klara SOK (2019) Blockchain technology and financial regulation: a risk-based approach to the regulation of ICOs. *Eur J Risk Regul* 10(2):263–314. <https://doi.org/10.1017/err.2019.41>
- DiMaggio PJ, Powell WW (1983) The iron cage revisited: Institutional isomorphism and collective rationality in organizational fields. *Am Sociol Rev* 48(2):147–160. <https://doi.org/10.1515/59780691229270-005>
- Eubanks V (2018) *Automating inequality: how high-tech tools profile, police, and punish the poor*. St. Martin's Press
- Fineman M (2021) Universality, vulnerability, and collective responsibility. *Les ateliers de l'éthique* 16(1):103–116. <https://doi.org/10.7202/1083648ar>
- Fineman MA (2008) The vulnerable subject: Anchoring equality in the human condition. In: *Transcending the boundaries of law*. Routledge-Cavendish, pp 177–191
- Gilley B (2017) Technocracy and democracy as spheres of justice in public policy. *Policy Sci* 50(1):9–22. <https://doi.org/10.1007/s11077-016-9260-2>
- Kim Y, Zhang J (2016) Digital government and wicked problems. *Gov Inform Quart* 33(4):769–776
- Klingbeil A, Grützner C, Schreck P (2024) Trust and reliance on AI—An experimental study on the extent and costs of overreliance on AI. *Comput Human Behav* 160:108352. <https://doi.org/10.1016/j.chb.2024.108352>
- Levin K, Cashore B, Bernstein S, Auld G (2012) Overcoming the tragedy of super wicked problems: constraining our future selves to ameliorate global climate change. *Policy Sci* 45(2):123–152. <https://doi.org/10.1007/s11077-012-9151-0>
- Miller PB, Gold AS (2015) Fiduciary governance. *William & Mary Law Rev* 57(2):513–586. <https://heinonline.org/HOL/Page?handle=hein.journals/wmlr57&id=525&collection=journals&index=>
- Prainsack B (2015) Three H's for health – the darker side of big data. *Bioethica Forum* 8(2):40–41. http://www.bioethica-forum.ch/content/d_AktAusgabe.php
- Rittel HW, Webber MM (1973) Dilemmas in a general theory of planning. *Policy Sci* 4(2):155–169. <https://doi.org/10.1007/BF01405730>
- Taffel S (2012) Escaping attention: digital media hardware, materiality and ecological cost. *Cult Mach*, 13
- Torcal M, Christmann P (2021) Responsiveness, performance and corruption: reasons for the decline of political trust. *Front Polit Sci* 3:676–672. <https://doi.org/10.3389/fpos.2021.676672>
- van de Poel I (2017) Society as a laboratory to experiment with new technologies. In: Bowman DM, Stokes E, Rip A (eds) *Embedding new technologies into society*. Jenny Stanford Publishing, pp 61–87
- van der Meer TWG, van Erkel PFA (2024) Moving beyond the political trust crisis debate: Residual analyses to understand trends in political trust. *Eur J Polit Res* 63(3):1240–1257. <https://doi.org/10.1111/1475-6765.126450.1111/1475-6765.1264510.1111/1475-6765.126450.1111/1475-6765.12645>
- Wegner D, Verschoore J (2022) Network governance in action: functions and practices to foster collaborative environments. *Adm Soc* 54(3):479–499. <https://doi.org/10.1177/00953997211024580>

- Winkel O (2021) The digitally induced increase of wicked problems as a challenge for politics and public management. *J Public Adm Policy Res* 13(1):1–10
- Yeow A, Faraj S (2014) Technology and sociomaterial performance. In: *Information Systems and Global Assemblages.(Re) Configuring Actors, Artefacts, Organizations: IFIP WG 8.2 Working Conference on Information Systems and Organizations, IS&O 2014, Auckland, New Zealand, December 11-12, 2014. Proceedings.* Springer Berlin, pp 48–65
- Zhang J, Kim Y (2016) Digital government and wicked problems: solution or problem? *Inform Polity* 21(3):215–221. <https://doi.org/10.3233/IP-160395>

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.



Part I
Governance: By and of Technology

Municipalities Enacting Algorithms: A Typology of Dutch Municipal Strategies for Leveraging Algorithmic Systems



Maranke Wieringa

Abstract Increasingly, the public sector and governments are algorithmizing. Inherently, the algorithms they use are sociotechnical in nature. More concretely, algorithms are done or rather: ‘enacted’. Similar or technically identical systems can be enacted differently by another organization which can completely shift the legal, ethical and social implications of their use. If academic efforts of providing legally, ethically, and socially acceptable ways of doing algorithms are to succeed, we urgently need to understand how algorithmic systems are enacted in practice. This contribution presents an analysis of municipal algorithmic practices in Dutch municipalities, based on a wide variety of ‘scavenged’ material such as fieldnotes, interviews, and documents and a four-year long engagement with the Dutch municipal scene. Departing from this analysis, the typology of different municipal enactment strategies is synthesized. Four archetypes are introduced: frontrunners, fast fashionistas, conscientious cat-observers, and slowpokes. These archetypes enact algorithms in different ways and with varying goals. Each of these archetypes has benefits and drawbacks, and recommendations are highly situational. However, for municipalities just setting out to grapple with this theme, a conscientious cat-observer strategy is potentially more compatible with current public and political demands around responsible and accountable algorithm use.

M. Wieringa (✉)

Parell Group, B.V., Arnhem, The Netherlands

Data School, Utrecht University, Utrecht, the Netherlands

Department of Media and Culture Studies, Utrecht University, Utrecht, the Netherlands

e-mail: maranke.wieringa@parell.nl

© The Author(s) 2025

J. Goossens et al. (eds.), *Public Governance and Emerging Technologies*,
https://doi.org/10.1007/978-3-031-84748-6_2

1 Introduction

On December 22nd, 2022, the State Secretary of Digitalization officially launched the national algorithm registry of the Netherlands.¹ All relevant algorithms, defined as ‘a set of rules and instructions that a computer follows automatically to make calculations in order to solve a problem or answer a question’, require registration in the coming years. The goal? To make algorithms ‘transparent’ and to ‘explain their application and outputs’, so that governmental algorithm use becomes ‘inspectable’ by citizens, the media, etc.² The installment of the national algorithm registry was spurred by upheaval about various kinds of harmful or potentially harmful algorithm use by Dutch governmental organizations in earlier years.³ In the wake of these incidents, Dutch governmental organizations started several initiatives to *do* algorithms *better*.

This raises the question as to how governmental organizations ‘do’ or rather *enact* these socio-technical algorithmic systems and enact them *well*.⁴ Here, I will focus specifically on Dutch municipalities: a large and pluriform group of low-level governmental organizations with varying levels of inhabitants, operational capacity, and algorithmic ambitions. Examining how Dutch municipalities enact algorithmic systems, responsibly and accountably manner, both in practice *and* on paper might then provide foundational knowledge on how to accountably and responsibly algorithmize the public sector.⁵

2 The Algorithm in the Room

As we grappled with the initial pandemic awkwardness of video calling and internet breakdowns, I interviewed a municipal data protection officer partially through Teams and partially through phone. They half-joked that if their colleagues knew that they were talking to me about algorithms that they’d be startled by the term and would say ‘Algorithms?! We don’t use any algorithms in our municipality?!’.⁶ This was amongst the most recurring sentences I heard spoken when I disclosed my research topic to civil servants during fieldwork.

I found this intriguing, as decades-old automation efforts left governments with a plethora of algorithmic systems. Yet algorithms, it seemed from interactions such as these, are Something Done Elsewhere by Other Organizations. Often, algorithms

¹ANP (2022).

²Digitale Overheid. (n.d.).

³For instance SyRI and the Childcare Benefits Scandal—see, for example, Wieringa (2022); Konaté and Pali (2023).

⁴See Seaver (2017) and Wieringa (2020).

⁵See Meijer et al. (2021); Meijer and Grimmikhuijsen (2020).

⁶T1.

are taken by civil servants to denote the ‘high tech’: AI, machine learning, neural networks, and the like. We see this play of words in the phrasings of respondents and documents released by the municipalities.⁷ ‘We do use decision trees and web forms that automatically route information, *but we do not use* automated decision-making in our municipality’.⁸ Simultaneously when asked to explain what algorithms are, the same respondent notes that algorithms are also present in the web forms they mentioned. Even within the same municipality we often do not find nor should we perhaps expect consistency in the definition of algorithms.⁹

Such inconsistencies highlight two slippery problems that need addressing. First, there is a focus on ‘new’ technologies. While the use of new/more accessible technological possibilities in public administration, such as machine learning, certainly warrants investigation,¹⁰ they are not the only algorithmic techniques that shape public administrations in fundamental ways.¹¹ The decades-long automation of public administration left us with an enormous amount of algorithmic systems, often in the shape of ‘simple’ decision trees or statistical analyses. As such we would expect a plethora of algorithms in use in governmental organizations. Yet, when prompted, many civil servants argue that they do not use algorithms. These ‘simple’ and/or legacy automated systems often go unrecognized, undetected, and unthought of. Or rather, they go unnoticed until things go wrong or turn out to have become controversial over time.¹² This complicates the present research. For we are not only trying to trace the adoption and how-to of AI within public administration, but we are also tracking the awareness and recognition of the continual use of algorithmic systems throughout entire organizations.

Second, while governments might embrace a broad definition of algorithms on paper, this does not mean that this is also how definitions are felt and practiced on the ground. Despite the ambitions of—amongst others - the national algorithm registry, a consistent understanding of algorithms within and across organizations is lacking. Thus, any formal presentation of the state of affairs will paint a meagre and skewed picture. The slippery nature of enacting algorithmic systems cannot be solely captured in formal documents. Yet, studying it solely in practice does not paint a complete picture either.

This chapter therefore aims to provide a combined, long-exposure view of paper *and* practice over 5 years (2018–2023) time. I investigate both the ‘new’ *and* the ‘old’ algorithmic systems. In doing so, I aim to sketch a more comprehensive view on algorithmic practices in public administration in general. Practically, this chapter is divided into three parts. First, I will describe my analytical approach. Secondly, I

⁷Field notes 25 May 2021; Field notes 30 November 2021; Field notes 19 November 2020; Field notes 18 March 2021; T1.

⁸T1.

⁹See, for example, Rekenkamer Metropool Amsterdam (2023) and Gemeente Rotterdam (2023).

¹⁰See, for example, Medaglia and Tangi (2022), Mikalef et al. (2019, 2022).

¹¹See, for example, Bovens and Zouridis (2002), van Eck (2018), Wieringa (2020, 2022), Zouridis et al. (2020) and Zuurmond (1994).

¹²See Wieringa (2022).

will discuss how algorithmic systems are enacted based on my observations and analyses. This part is structured by communication interrogatives (e.g. what/why/who?).¹³ Finally, I zoom out to sketch archetypical municipal enactment strategies, which are distilled from the analysis.

3 Analytical Approach

With the terminology of enacting algorithmic systems and/or algorithmic accountability, I invoke a sociomaterial perspective. Following Mol, Seaver, and Orlikowski, I take algorithms not to be solely technological artifacts. Rather, I take them as ‘enacted’ sociotechnical systems.¹⁴ They come into being at the intersection of algorithmic technology, data, social practices, and (organizational) culture. As Goffey notes, ‘[a]lgorithms obviously do not execute their actions in a void’.¹⁵ They can only ‘work’ when they are fed data, implemented in work practices, and have some kind of effect.

By looking at the sociotechnical algorithmic system, instead of a sole focus on algorithms, I required different material than just observations of code in the wild. Rather than studying recipes or step-by-step instructions, I investigate how these recipes are thought of, designed, leveraged, and how they effect change as a result of humans, technology, policy, and (organizational) structures ‘intra-acting’.¹⁶ That is, I take a sociomaterial stance in which I subscribe to the view that an absolute separation between technology/materiality, social practices, and culture is impossible and does not do justice to their capacity to affect change.¹⁷

3.1 Data Collection

In this chapter, I thus draw on heterogeneous data to examine how algorithmic systems are enacted, on paper and in practice. For the practice-part I draw on scavenged materials.¹⁸ These materials include:

- Field observations from 86 encounters I have had with civil servants between September 2018 and November 2022.¹⁹

¹³ cf. DAMA International (2017).

¹⁴ Mol (2002), Seaver (2017), Orlikowski (2007) and Wieringa (2020).

¹⁵ Goffey (2008), p. 18.

¹⁶ Barad (2007), p. 141.

¹⁷ Orlikowski (2007).

¹⁸ See Seaver (2017).

¹⁹ Interactions with my respondents included observing project meetings, observing impact assessments or inter-municipal meetups about algorithms, participating in working groups and advising

- 3 semi-structured interviews with civil servants actively working on responsible algorithm use between July and August 2020;²⁰
- 21 documents collected in the field or sent to me by respondents between September 2018 and November 2022.²¹
- All 104 municipal entries in the national algorithm registry (across 8 municipalities);²²
- 83 documents from public council information systems of 47 municipalities.²³

The first three strands of data collection relied on a combination of purposive sampling, convenience sampling and snowball sampling. In order to get in touch with otherwise hard-to-identify civil servants, I, for example, often attended inter-municipal data/algorithm meetups where I visibly identified myself as a researcher and invited conversation with interested attendees. This strategy relied on self-selection amongst a purposively selected group of individuals. This often led to snowball interactions later on (e.g. being invited to observe project meetings in a specific municipality).

However, this strategy did muddy clarity of scope in the sense of absolute numbers of included municipalities. Many of my interactions pertain to intermunicipal meetings in which it often was not (exhaustively) clear which municipalities participated—especially when participants did not actively join group discussions or posed questions. What can be said is that the entire material spans *at least* across 4/4 G4 municipalities,²⁴ 25/41 G40 municipalities,²⁵ 11/299 remaining municipalities and 12 municipal-transcending organizations (e.g. the Association of Dutch Municipalities, or regional collaborations between municipalities).²⁶

governments about guidelines or fairness assessments at their request. In this sense, I have also been an active agent in the field I have been studying.

²⁰ Labeled as T1, T2, T3.

²¹ Emerson et al. (2001), p. 356.

²² This reflects the state of affairs on 19 September 2023.

²³ The 45 largest municipalities in the Netherlands were included in this query, and were supplemented with any smaller municipalities that had registered algorithms in the registry at that time ($n = 2$). The council information systems of the selected municipalities were queried for information about [algoritme*] between the period of 1-9-2018 to 18-9-2023. In several systems it was also not possible to set a specific to-from date in the search function. In these cases the years 2018–2023 were used and the cut-off was applied manually. This resulted in a total of 83 documents, amongst which 42 were found in the four largest municipalities (G4) and 41 were unevenly distributed across the remaining 41 municipalities.

²⁴ The G4 are the four largest municipalities in the Netherlands.

²⁵ The G40 are the 41 largest municipalities in the Netherlands after the G4.

²⁶ This reflected the number of municipalities in November 2023.

3.2 Methodology

The methodological approach taken in the analysis of the collected data is indebted to Grounded Theory (GT). Yet whereas GT is inductive, the present analysis supplements GT's inductive approach with a deductive approach in the abstraction process. In a nutshell, GT departs from empirical data from which theory eventually emerges. The present research takes a similar stance but in the process of generalizing theory, it supplements the inductive process with deductive reasoning in that, when applicable, developed theory is used to guide the abstraction process (e.g. frequent comments and observations about public value, operational capacity, and the authorizing environment could be linked to Moore's strategic triangle).²⁷ In doing so, the inductive approach is strengthened and made more generalizable without sacrificing its bottom-up, emergent qualities. That is, where applicable existing theory can be used to describe and support the inductive results and abstractions, but only when this fits with the data (see Fig. 1).

The scavenged data (notes from field observations, interviews, documents obtained during fieldwork, and documents extracted from council information systems) were coded line-by-line in NVivo in two phases. Open coding was used to

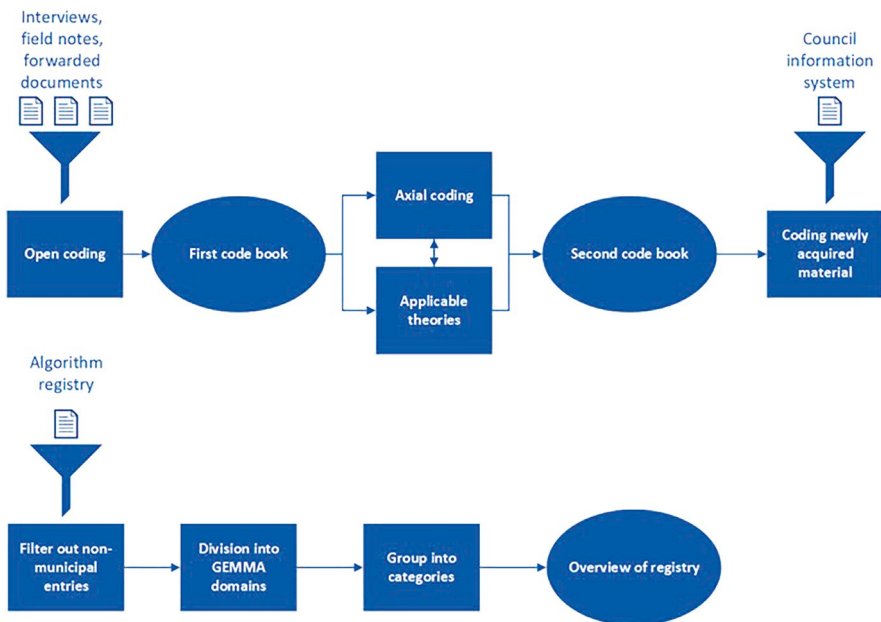


Fig. 1 Coding process and analytical steps

²⁷ Moore (1995, 2013).

identify the relevant themes for subsequent categorization.²⁸ Secondly, these open codes were ordered, categorized and used to create new, more abstract, axial codes, which shifts the focus from discrete observations to more abstract linkages or ‘axes’ between these observations. Normally, GT involves a third and final round of coding which focuses on these central categories to then ‘arrive at concepts that help to explain the phenomenon’.²⁹ In combining the inductive and deductive, however, such an abstractive move is already incorporated in the research design, and saturation was achieved after the second round of coding and deductive supplementation.

Additionally, in order to give insight in the breadth of algorithmic applications, I also draw on the national algorithm registry. A complete overview of the national algorithm registry was downloaded on September 19th 2023. Only data about municipal algorithms was selected from this (104 municipal algorithms across 8 municipalities). The municipal algorithms were then divided into domains based on the Dutch model architecture for municipalities (GEMMA).³⁰ The registrations were subsequently grouped by application type (e.g. resource/capacity prediction, enforcement; Fig. 1).

4 Analysis

In the remainder of this chapter I sketch how algorithmic systems are being enacted. The analysis is structured by breaking down the overarching question of ‘how algorithmic systems are enacted into its different elements, namely:

- Who is involved in enacting algorithmic systems?
- What enactments can be identified?
- Where are algorithms being enacted?
- How quickly and systematically are algorithms being enacted?
- Why and when are algorithmic systems being enacted?
- Whom does it affect?

Below, I will highlight each of these aspects in more detail. After this, I will discuss the four archetypical strategies which surfaced from the analysis in the second part of this chapter.

²⁸ Williams and Moser (2019), p. 48.

²⁹ Denscombe (2017), p. 116.

³⁰ GEMMA (Gemeentelijke Model Architectuur, tr. Municipal Model Architecture) is the reference architecture used for Dutch municipalities. It is used as a blueprint for the information service of municipalities. It describes the following domains: social domain, physical environment, public services, public order and safety, management, and the supporting domain. See <https://www.gem-maonline.nl/>.

4.1 *Who Is Involved in Enacting Algorithmic Systems?*

Algorithms are mostly recognized in municipalities once they start working with more complicated analyses. This often coincides with the presence of data labs or a data-driven transition within the organization. As highlighted before, this does not mean that algorithms are not present in municipalities prior. Yet such ‘simple’ algorithms are generally rarely perceived or noticed.

This greatly complicates the question of ‘who is involved in enacting algorithmic systems’. We do not have numbers for ‘algorithm work’ in the Dutch public sector. For whereas a few, large, municipalities have dedicated roles for ‘algorithm work’ (e.g. ‘AI officer’, ‘algorithm expert’, ‘advanced analytics’, ‘robotic process automatization’), this is a rare exception. For the remaining municipalities, ‘algorithm work’ is not necessarily clearly labelled. What we can draw on for this group, are insights in the data labs which generally are amongst the first places where municipalities start to perceive algorithms. Once an organization matures in their data analytics competency, algorithm-related questions often pop up.³¹ Many respondents from small municipalities subscribed to the necessity of thinking about algorithm-related issues, but noted that their organization first needed to grapple with data before they could turn their attention to algorithmic systems. ‘Data work’ is thus seen as a precursor to ‘algorithm work’.

In general, municipal data or algorithmic projects involve a multidisciplinary project team of human actants. The precise make-up of a municipal team working with algorithmic systems differs per situation. As a departure point, there are some common roles that are identified by VNG Realisatie’s inventory of municipal data labs.³² Paramount roles are those of a project leader and data specialists (e.g. data engineers, data scientists and/or data advisors). Other team members could be for instance a privacy officer, information/data manager and any kind of advisors (e.g. policy advisor, strategy advisor). In some cases, municipalities actively involve street-level bureaucrats whose daily operation is influenced by the introduction of an algorithmic system.³³ There may also be a collaboration between the municipality and an external party.

Though the data of VNG Realisatie does not note the size of the municipality, based on the field work, recurring remarks and field observations signal a connection between the size of the municipality and the operational capacity devoted to data- and algorithmic work. A respondent from an executive organization working for three small municipalities, illustrates: ‘Amsterdam has a whole department. We are supposed to do it on the side’.³⁴ This also connects to the existence of dedicated algorithm-roles in such large (especially ‘G4’) municipalities.

³¹ See Rekenkamer Rotterdam (2021).

³² See VNG 2019.

³³ Field notes 12 December 2018; 15 May 2019; 10 February 2020.

³⁴ Field notes 14 October 2019.

This operational capacity gap is translated in expectations and dependencies between municipalities. As smaller municipalities have less means (manpower and budget) to invest in the design and exploration of algorithmic systems, they look to large municipalities to take the lead.³⁵ Larger municipalities are expected to chart the algorithmic landscape and develop best practices which smaller municipalities can adopt.³⁶

4.2 What Enactments Can Be Identified?

There is a certain, often implicit, hierarchy at play which very much relates to *what* kind of algorithms are enacted. In part, this hierarchy is reified in the Guidelines for algorithms of the Ministry of the Interior and the Ministry of Justice and Security to which many municipalities defer.³⁷ In their guidelines—which are fundamental touch stones for further innovation and usage of algorithmic systems by governments—the Ministries note that:

We can differentiate between four ‘ways of using algorithms’:

1. *Descriptive*

An analysis of ‘what is happening’—the algorithm gives a rendition of what it is perceived.

2. *Diagnostic*

Analysis of ‘why is it happening’—the algorithm gives a probability or an analyses of why something happens, guided by positive/negative or predictive values, or calculates a probability as input for work processes and/or decision making.

3. *Predictive*

Analysis of ‘what could happen’—the algorithm sketches an expectation, a prediction of what could happen, or a chance/probability of a future event.

4. *Prescriptive*

Analysis of ‘what should happen’—the algorithm decides and/or dictates the decision/ action or execution.

The idea and logic of this hierarchy is that ‘descriptive’ and ‘diagnostic’ are perceived as having a low impact. They look back in time and describe/explain, but they do not look forward and influence the state of affairs. This is different, it is argued, for predictive and prescriptive uses. Here, algorithmic systems look towards the future and may influence the state of affairs. Forward-looking algorithmic systems are assumed to be more impactful as they directly or indirectly influence the state of affairs due to their recommendations, predictions, and/or automatic decision making. That is, the more autonomy an algorithmic system is granted, the more

³⁵The G4 denotes the 4 largest municipalities in the Netherlands. The G40 the 41 largest municipalities after these four.

³⁶T1.

³⁷See Ministerie van Justitie en Veiligheid (2021).

scrutiny it seems to deserve.³⁸ This ties in with the more complex analyses leading to more (perceived) algorithm use, as described in the previous section.

Additionally, the guidelines make note of the type of technology that is leveraged and its inherent complexity and the degree to which people can understand its workings. A ‘simple’ decision-tree is deemed uncomplex and very explainable. An IF this, THEN that logic is very understandable, it is posited. Deep learning models, in turn, are deemed highly complex and difficult to explain and understand. Explaining to a lay person the many hidden layers of a deep learning model and their workings is a much bigger challenge, it is implied.

4.3 Where Are Algorithms Being Enacted?

Algorithmic systems are not reserved for one single municipal domain. It is not possible to provide an exhaustive list of applications, because inventorying has either not started or is ongoing.³⁹ Yet, it is possible to provide examples of algorithmic systems used in all municipal domains. Below I will sketch algorithmic endeavors of municipalities across these different domains: social, physical environment, public services, public order and safety, management, and support. An overview and examples of the algorithms present in the national algorithm registry per domain and type can be found in Table 1 (n = 104 across 8 municipalities).

Though the algorithm registry, shows the breadth of application areas, it paints far from a complete picture as registration is not compulsory. A comparison between the council information documents and the algorithm registry revealed that 18/47 municipalities said that they make use of algorithmic systems while not having any entries in the registry. Moreover, even municipalities that have committed to registering their algorithmic systems struggle to do so. Thus, their contribution to the registry is as of yet incomplete. Evaluations of inventory efforts note that it proves very difficult in practice, for instance because algorithmic systems go unrecognized.⁴⁰

4.4 How Quickly and Systematically Are Algorithms Being Enacted?

It is thus difficult to pin down the unnoticed and unthought of algorithmic systems. This also complicates sketching a complete picture of how these are being enacted. This is important, as not every municipality adopts algorithmic systems in the same

³⁸ Ibid.

³⁹ E.g. Rekenkamer Metropool Amsterdam (2023); Gemeente Rotterdam (2023).

⁴⁰ See Rekenkamer Metropool Amsterdam (2023); Gemeente Rotterdam (2023).

Table 1 Overview of national algorithm registry per September 19th 2023

Domain	Type of algorithmic system	Number of registrations	Number of different municipalities
Social domain	Resource/capacity prediction	4	2
	Preventative/curative	4	3
	Vetting eligibility	8	5
	Fraud prevention/welfare benefits re-evaluation	3	3
	Monitoring	2	1
Physical environment	Enforcement	19	6
	Traffic management	12	3
	Waste management	6	3
	City planning	5	3
	Water management	3	2
	Maintenance and end of life monitoring	2	1
	Air quality management	1	1
	Noise and sound management	1	1
	Fauna management	1	1
Public services	Estimating ‘WOZ’ worth of buildings	3	3
	Estimating viability of companies	1	1
	Making documents accessible	1	1
	Democratic process	2	2
Public order and safety	Analysis	6	2
	Enforcement	3	2
	Preventative/curative	7	3
Management	Process optimisation	1	1
Supporting domain	Anonymization	3	3
	Analysis	2	2
	Routing	1	1
	Data quality control	1	1
	Making documents accessible	1	1
	Determining fees	1	1

way. For instance, some make a conscious, strategic, decision to adopt more algorithmic systems throughout their organization. Others may make more ad hoc calls. In other words, the municipal behavior around algorithmic systems varies. Three aspects influence how and when such algorithmic systems are enacted: the speed of recognition and/or adoption and the structural engagement/embedding thereof, and the required awareness/knowledge/expertise/governance in the organization to ensure responsible and accountable algorithm use. This latter point often requires

organizations to adapt and accommodate algorithmic practices, a phenomenon termed ‘algorithmization’.⁴¹

This then brings us to how sustained or embedded algorithm adoption in municipalities is. There are stark differences of organizational embeddedness of algorithm use. Some municipalities use algorithmic systems on an ad hoc basis in one or a few use-cases. Other municipalities draw on algorithmic systems strategically throughout all or most of their endeavors. The latter is mostly found within organizations which have a higher level of operational capacity and awareness (more on this below). Yet in order to strategically use algorithmic systems, organizations also need to pay high-level attention to how, when and when not to use these systems. In ad hoc projects such organization-wide engagement is not per se required. On the other hand, there are also those municipalities who use a one-off project to *also* kickstart this high-level positioning. As seen above, some frontrunner municipalities, typically G4 or large G40 municipalities, advocate for the ‘eat your own dog-food’ approach. They structurally reserve capacity in the organization to develop and retain specialized knowledge. This approach finds its solution in bringing expertise *into* the organization.

Enacting algorithms in a public sector organization also requires new or supplemented awareness, expertise and knowledge. Such awareness/expertise/knowledge is required to recognize algorithms in practice. Yet, respondents and reports often note that there is an organizational lack of knowledge and awareness.⁴² As one civil servant puts it: ‘there is zero knowledge about algorithms [in our municipality]’.⁴³ In organizations where there is some level of awareness and expertise, this may not be evenly distributed. Recurring remarks signal, ‘an enormous rift between a digitally-skilled and the digitally-unskilled government. A rift that exists at all levels’.⁴⁴

Yet upping expertise, knowledge, and awareness about algorithmic systems is not as simple as ‘just add more people’. As a respondent notes the bottleneck—when dealing with algorithms and data—is not so much the technical side of things, ‘but rather the culture, skills, knowledge and attitudes’.⁴⁵ This again highlights the importance of seeing algorithms not as technical objects in isolation, but as socio-technical systems which involve infrastructures, the organization, project teams, work routines, or in short all kinds of technical, material, social, and cultural aspects. If we add more experts, we do not necessarily change the rest of this equation, thus potentially leading to little change.

⁴¹ See Meijer et al. (2021); Meijer and Grimmelikhuijsen (2020).

⁴² E.g. Rekenkamer Metropool Amsterdam (2023); Gemeente Rotterdam (2023); T1; T2; Field notes 30 April 2020; Rekenkamer Rotterdam (2021).

⁴³ Field notes 23 March 2021.

⁴⁴ Field notes 17 March 2021; see also field notes 28 October 2021.

⁴⁵ Field notes 25 February 2019.

4.5 *Why and When Are Algorithmic Systems Being Enacted?*

The precise need for a data-driven project and the approach thereof is very situational. Municipalities often hope to gain ‘better results with less resources’.⁴⁶ Algorithmic projects in the social domain, particularly for welfare benefit fraud risk detection, seem enticing. They are one of the few so-called, ‘business cases’ that municipalities have when it comes to the social domain. As such systems hold the promise of generating income (by revoking welfare benefits and retrieving wrongfully paid benefits). Their financial success is one of the evaluation metrics. One municipality noted for instance that ‘the system paid for itself’.⁴⁷ Efficiency and efficacy can thus be strong drivers for the adoption of algorithmic systems.

Yet efficiency and efficacy are not the only public values at stake.⁴⁸ There is increasing attention to what the WRR terms ‘anchoring principles’ such as privacy and autonomy.⁴⁹ This paradigm shift in perception about public values at play is beautifully illustrated by a Mentimeter poll at one event. The organizers asked the participants to write down which public values were the most important when working with data. Several answers popped up on the screen without discussion: transparency, privacy, trustworthiness, integrity, equality. Then efficiency appeared. The host remarked: ‘Someone dared to write down efficiency. Or well ... dare ...’.⁵⁰ The implication of such a remark is that a, sole, focus on efficiency is out of the ordinary, perhaps even frowned upon.

Today, many documents and respondents acknowledge the importance of safeguarding public values in algorithm use.⁵¹ In part incidents played a large role in this: many respondents refer to the Dutch incidents of the SyRI case and the child-care benefits scandal.⁵² Spurred by these incidents or not, ethics and a discussion about good conduct and fundamental rights has gained momentum.⁵³ In light of this, many municipalities that are consciously adopting and/or recognizing AI/algorithmic systems now implement safeguards and checks and balances to enact algorithmic systems well.⁵⁴ There are also municipalities that take the lessons and best practices from this former group to implement their own algorithmic practices and safeguards. In many cases, the algorithm registry is one of several possible

⁴⁶ Field notes 19 October 2020.

⁴⁷ Field notes 10 February 2020.

⁴⁸ See Moore (1995, 2013).

⁴⁹ See Wetenschappelijke Raad voor het Regeringsbeleid (WRR) (2011).

⁵⁰ Field notes 28 October 2021.

⁵¹ See Ministerie van Binnenlandse Zaken en Koninkrijksrelaties (2021); Ministerie van Economische Zaken en Klimaat (2018); Overheidsbreed Beleidsoverleg Digitale Overheid (2020); Ministerie van Economische Zaken en Klimaat (2019); Raad voor het Openbaar Bestuur (2021).

⁵² T1; T2; Field notes 31 March 2020 and 20 May 2020.

⁵³ T1; Field notes 18 September 2019; 6 February 2020; 26 February 2020; 18 March 2020.

⁵⁴ Field notes 22 March 2021; Rekenkamer Metropool Amsterdam (2023); Gemeente Rotterdam (2023); Gemeente Den Haag (2020); Gemeente Utrecht (n.d.).

safeguards. Other, possible or realized, mitigating measures include, amongst others, a fundamental rights impact assessment, procurement guidelines, policy documents, audits, and technical measures such as bias testing. Some of the larger municipalities have explicitly adopted an ‘algorithm governance’ which stipulate which measures need to be coupled to ‘high risk’ algorithmic applications.⁵⁵ Other municipalities adopt new algorithmic systems quickly, in order to capitalize on their technological promises, but fail or are felt to lack coupling this with appropriate checks and balances.⁵⁶

Another thing that these incidents spurred is fear for algorithms.⁵⁷ Subconsciously, it seems, civil servants differentiate between ‘algorithms’ and ‘Algorithms’. Automated decision-making, AI, image-recognition, and predictive systems are implicitly seen as the exotic, or even mysterious and/or dangerous,⁵⁸ ‘Other’ of the algorithm-world.

Algorithms, with a capital A, are—to civil servants—those computational systems that are deeply complicated, and are perceived to exist materially and/or symbolically out of their immediate surroundings. Equating Algorithms to algorithms is perhaps also somewhat of a way to avoid societal upheaval or to ease colleagues into the theme. One respondent continues:

The word ‘algorithm’ is a bit scary, I notice. I think that when you talk about flowcharts the fear is lessened. I also reason in this way. Algorithms have a bit of a sour aftertaste, especially since SyRI. When I talk about algorithms then some people say: ‘Algorithms? No, we do not have any algorithms in [their municipality]’. Though, when we see them as these step-by-step descriptions, then we do have them, of course. Municipalities are a bit nervous about it.⁵⁹

Algorithms, with a capital A, are the things some municipalities do not seem to want to touch with a ten-foot-pole if it can be helped. Though, as we have seen in the previous section, there are stark differences here.

4.6 *Whom Does It Affect?*

Even within a single municipality, the ‘Algorithm’-fear sketched above is not felt evenly across all municipal algorithmic systems. Much of this has to do with whom is affected by the deployment of an algorithmic system. Often a differentiation is made between the social domain and the physical domain. The physical environment domain of a municipality generally has a long history of working with data and algorithmic systems: geodata and automated traffic management systems are

⁵⁵ See Rekenkamer Metropool Amsterdam (2023) and Gemeente Rotterdam (2023).

⁵⁶ Field notes 20 May 2020.

⁵⁷ Field notes 29 October 2020.

⁵⁸ Field notes 17 December 2018.

⁵⁹ T1.

but two examples thereof. But more importantly, aside from the physical domain having more experience, as one respondent summarizes: ‘the physical domain is not that difficult, but the social domain is much more vulnerable’.⁶⁰

Applications in the social domain are generally felt to be much more intrusive and potentially infringing on human rights and encroaching on anchoring values and are felt by many respondents to require special attention for instance in the guise of ethical discussions.⁶¹ To illustrate, a system regulating traffic flow, does not impact citizens as much as a system scrutinizing potential fraud with welfare benefits, the reasoning goes. That is, a decision to make you wait a minute longer in front of a red light has less potential to harm than a decision which may upend your finances.

As each algorithmic system involves specific stakeholders, a more detailed analysis of whom is potentially affected is not possible at this macro-level. Yet, the general tendency is that social domain applications are more likely to impact individuals, as this domain exists mainly to support citizens with care, finances, and aid.

5 Archetypical Strategies of Algorithm Enactment

In the remainder of this chapter, a synthesis of archetypical positions will be introduced based on the identified key behavioral elements (Table 2) as well as field observations of municipalities. Four archetypical positions can be identified: frontrunners, fast fashionistas, conscientious cat-observers, and laggards (Table 3). Frontrunners are municipalities that were quick to adopt and recognize algorithmic systems. They invest in algorithmic projects and responsible and accountable algorithmization. Fast fashionistas are municipalities that were quick to adopt

Table 2. Key behavioral elements

Who	• Dedicated role for ‘algorithm work’ (e.g. ‘AI officer’, ‘algorithm expert’) • Dedicated team/s for data analysis
What	• The different kinds of algorithm usage of the organization
Where	• Domains involved • Registrations in register
How	• Ad-hoc/organization-wide algorithm usage • Existence of algorithm policy/governance documents (proactive vs. reactive) • Degree of attention to socio-cultural questions (e.g. awareness)
Why	• Focus on efficacy and efficiency or focusing on a broader set of public values • Reasons for not consciously starting algorithm work: fear and/or data immaturity
When	• Self-professed quick or slow to adopt
Whom	• Considerations of impacts on stakeholders

⁶⁰ Field notes 20 November 2018.

⁶¹ E.g. field notes 18 September 2019; 1 October 2019; 3 October 2019; 8 February 2022; 9 February 2022 and 17 February 2022.

Table 3 Archetypical strategies of responsible and accountable algorithm enactment

	Frontrunners	Fast fashionistas	Conscientious cat-observers	Slowpokes
Description	<i>Municipalities that were quick to adopt and recognize algorithmic systems. They invest in algorithmic projects and responsible and accountable algorithmization.</i>	<i>Municipalities that were quick to adopt algorithmic systems, but spent little or no time and resources on reflecting on and ensuring responsible and accountable algorithmization.</i>	<i>Municipalities that are intentionally slower to adopt algorithmic systems and strive for high levels of responsible and accountable algorithmization.</i>	<i>Laggard municipalities with little to no conscious use of algorithmic systems or responsible and accountable algorithmization.</i>
Accountability	Ex ante: virtue/felt accountability. Ex post: compliance	Ex ante: none. Ex post: felt accountability/compliance	Ex ante: virtue/felt accountability. Ex post: compliance	None
Typically observable algorithm work				
Who	• Dedicated role for responsible/accountable 'algorithm work' • In-house expertise	• No dedicated role for responsible/accountable 'algorithm work' • Generally little in-house expertise • Technical reliance on contractors	• Dedicated team exploring responsible/accountable 'algorithm work' • Often a technical reliance on contractors	• No dedicated role for responsible/accountable 'algorithm work'
What	• A recognized, wide variety of systems, including forward-looking and those of significant technical complexity	• One or several predictive/prescriptive systems	• One or several predictive/prescriptive systems	• Not used
Where	• Numerous entries in the registry • Used across all domains	• No entries in the registry • Used in one or two domain	• A few to no entries in the registry • Used in one or two domain	• No entries in the registry • Not used

(continued)

Table 3 (continued)

	Frontrunners	Fast fashionistas	Conscientious cat-observers	Slowpokes
How	• Several policy/governance documents related to algorithms • Pro-active organization-wide guidelines for proper engagement with algorithmic systems • Awareness efforts under way • Implemented a variety of mitigating measures	• Discussion of algorithms in documents is generally limited to one or several ad hoc project/s but is not discussed across the organization • Discussion of responsible and accountable algorithmic conduct is reactive. • Little to no awareness • Little to no mitigating measures ex ante, mitigating measures may be added ex post	• Some documents in which algorithms are addressed and discussed • Guidelines are developed ad hoc based on best practices and lessons elsewhere, and evaluated before being implemented organization wide and are of a pro-active nature • Little awareness, but this is actively acknowledged • Implemented and/or experimenting with a variety of mitigating measures	• Little to no mentioning of algorithms in their documents • Little to no conscious engagement with the topic • No mitigating measures
Why	• Centering a variety of public values	• Centering efficiency and efficacy	• Centering a variety of public values	• Non-use often due to fear or low data maturity
When	• Quick adopter	• Quick adopter	• Slow adopter	• Non-adopter
Whom	• Ex ante mitigating measures such as a fundamental rights impact assessment	• No ex ante measures, possible ex post measures after potential upheaval	• Ex ante mitigating measures such as a fundamental rights impact assessment	• No mitigating measures

algorithmic systems, but spent little or no time and resources on reflecting on and ensuring responsible and accountable algorithmization ex ante. Conscientious cat-observers are intentionally slower to adopt algorithmic systems and strive for high levels of responsible and accountable algorithmization. Slowpokes, finally, are municipalities with little to no conscious use of algorithmic systems or responsible and accountable algorithmization.

These archetypes relate to algorithmic accountability in different ways. Some engage with a certain topic to avoid problems and scandals or because this is seemingly ‘the right thing to do’ (accountability as a virtue), others because they are/will

be required to do so (accountability as a mechanism).⁶² Simply put: some respond to the carrot, others wait for the stick.⁶³

At its core this boils down to different reactions to the anticipation of accountability.⁶⁴ As accountability *might*, or rather *will*, become a reality, actants alter their behavior accordingly.⁶⁵ This occurs both on the individual level as we saw above in light of felt accountability, and the organizational level.

We thus see that some municipalities anticipate algorithmic accountability in a proactive way. They either try to come up, or co-design certain standards—as happened with the algorithm registry—or they start exploring avenues to do algorithms in a more accountable way in their organization. Other municipalities, make a different assessment. For them, accountability may only be anticipated when it is required by law and regulation (compliance/accountability as a mechanism). Yet, anticipatory or felt accountability is experienced differently at various points in time. One civil servant notes: ‘algorithms were “cool” and “young”, no-one discussed the downsides back then’.⁶⁶ Due to the societal upheaval the fast fashionista municipality later faced, they were looking for a ‘stamp of approval’ in some form as a way to anticipate pressing questions from various corners.⁶⁷ Thus, some only start anticipating accountability once shit has hit the proverbial fan.

These archetypes will be explained below in more detail and an overview can be found in Table 3. What is important to note, is that these are archetypes, meaning they describe archetypical behavior. The strategies do not describe hard categories, rather it is a spectrum where municipalities tend to gravitate toward a certain strategy. Moreover, it is important to stress that the technical design and capabilities of algorithmic systems do not (solely) determine an organisation’s archetypical affiliation. I have, for instance, observed two municipalities who bought the same algorithmic system from a supplier. One of these municipalities gravitated towards a fast fashionista strategy, the other to a conscientious cat-observer strategy. The differences derived mainly from the way in which the system was enacted (e.g. meaningful human control, relation to the main process, testing, etc.). This further emphasizes the sociomaterial nature of doing algorithms.

⁶² See Bovens (2010).

⁶³ Field notes 18 May 2022, Meetup Algoritme- en sensorenregister.

⁶⁴ See Friedrich (1963) and Overman & Schillemans (2021).

⁶⁵ See Meijer (2005).

⁶⁶ Field notes 20 May 2020.

⁶⁷ Ibid.

5.1 *Frontrunners*

Frontrunner municipalities, such as Amsterdam and Rotterdam, have invested heavily in algorithmic practices. These municipalities tend to be large and tend to have the bigger budgets reserved for data-driven work and algorithmic practice. Such municipalities also tend to invest in developing practical tools which facilitate algorithmic accountability.⁶⁸ They very much anticipate algorithmic accountability and often depart from an accountability-as-a-virtue position.⁶⁹ Accountability for algorithms is ‘the right thing to do’. For instance, a coalition of frontrunner municipalities is working on several control instruments such a procurement guidelines, a governance model, and an algorithm registry. This kind of leading role with regards to the use of algorithmic systems and algorithmic accountability therefore is also expected of these frontrunners due to their larger capacity and budget. As said, they are expected to function as pioneers for the benefit of other, smaller municipalities.⁷⁰ The added benefit of this position is that through their anticipatory position, these municipalities may help co-design accountability standards and best practices.

5.2 *Fast Fashionistas*

Fast fashionistas, such as Nissewaard, are generally early adopters who were enthusiastic about the technological promises of algorithmic systems. They tend to be technological solutionist minded. Often, such municipalities, bought algorithmic systems from a third party or contracted them to design it for them when they were presented as ‘the next big thing’. Generally, such municipalities initially had little expertise and/or knowledge of these systems. Over time, however, such municipalities could encounter substantial problems because they may inadvertently have implemented insufficient accountability measures, their systems are deemed too opaque, they themselves do not know how their system works, the enacted algorithmic multiples are particularly disjointed, and/or because they inadvertently have taken insufficient risk mitigations and the like. The limited algorithmic literacy and awareness may thus impede initial anticipatory algorithmic accountability of this archetype. Given the current societal and political demands of accountability and responsible algorithm use, this archetypical strategy may be less preferable for public administrations.

⁶⁸ See Wieringa (2020).

⁶⁹ See Bovens (2010).

⁷⁰ E.g. T1.

5.3 *Conscientious Cat-Observers*

Another type of municipality decided to see which way the cat jumps prior to joining the algorithmic fray. An example is Amersfoort. These ‘conscientious cat-observers’ take things slow and made a purposeful decision to not be in the front group, but rather learn from best practices and mistakes from their peers. Such a municipality tends to carefully consider how to best deal with algorithmic systems, and often reserves capacity for organizational reflection. That is, they do not solely reflect on ‘is the system effective?’, a question shared by fast fashionistas, but also questions like ‘do we want to work this way, and if so, what kind of mitigation and checks and balances are required?’. While frontrunners innovate and develop control and accountability mechanisms ‘on the fly’ and fast fashionistas hurry to keep up with technological innovation, conscientious cat-observers purposefully build reflection and contemplation into their process. They are mindful of the risks involved, a sign of anticipatory accountability, and tread carefully, adopting proven best practices. Given the societal and political demands of accountability and responsible algorithm use, this archetypical strategy may be less preferable for public administrations who are starting with or considering ‘algorithm work’.

5.4 *Slowpokes*

Slowpokes are, as some would call them ‘laggards’. These municipalities have either just begun the data-driven transition or have not even considered this transition. Often these are small(er) municipalities that lack budget, expertise and capacity to make up the leeway. Due to the predominant manner of finding and connecting with respondents (e.g. data/algorithm meetups) this group is likely quite underrepresented in the data, as such events are more likely to be attended by municipalities that are more involved with algorithmic systems and able to devote man hours to such events. However, respondents’ recurrent comments about this group and their absence in the registry suggest that it concerns many municipalities.

6 Conclusion and Recommendations

We started this chapter with the question how municipalities enact algorithmic systems. In the analysis we have touched upon different elements that make up this enactment. Every element turned out to be ambiguous and pluriform and difficult to capture. Even more problematically, there are many known unknowns, as algorithms may not always be recognized as such. Harking back to the central question of how municipalities enact algorithms, the short answer is: ‘it depends’. However, four archetypical strategies were synthesized from the analysis and observations.

These four archetypical strategies that are identified are the frontrunners, fast fashionistas, conscientious cat-observers, and slowpokes (Table 3).

Each of these archetypes has general benefits and drawbacks. Yet, given the current public and political demands for responsible and accountable use of algorithmic systems, a fast fashionista strategy may prove riskier and thus less desirable for public administration. If a slowpoke organizations aims to join the algorithmic fray, a conscientious cat-observer strategy might fit them best. A conscientious cat-observer strategy minimizes risks and draws on best practices, making it potentially more compatible with current public and political demands around algorithm use. A frontrunner strategy requires substantial effort and operational capacity and may not be realistically achievable for most municipalities.

References

- Algoritmeregister voor de overheid. Algoritmeregister voor de overheid. Retrieved March 4, 2024, from <https://www.digitaleoverheid.nl/overzicht-van-alle-onderwerpen/algoritmes/algoritmeregister/>
- ANP (2022, December 21) Kabinet publiceert register met algoritmes van de overheid. Leidsch Dagblad. <https://www.leidschdagblad.nl/binnenland/kabinet-publiceert-register-met-algoritmes-van-de-overheid/11486570.html>
- Barad K (2007) Meeting the universe halfway: quantum physics and the entanglement of matter and meaning. Duke University Press
- Bovens M (2010) Two concepts of accountability: accountability as a virtue and as a mechanism. West Eur Polit 33(5):946–967. <https://doi.org/10.1080/01402382.2010.486119>
- Bovens M, Zouridis S (2002) From to street-level to system-level bureaucracies: how information and communication technology is transforming administrative discretion and constitutional control. Public Adm Rev 62(2):174–184
- DAMA International (2017) *DAMA-DMBoK: data management body of knowledge*, 2nd edn. Technics Publications
- Denscombe M (2017) The good research guide: for small-scale social science research projects, 6th edn. McGraw-Hill Education
- Digitale Overheid (n.d.) Werkagenda. <https://www.digitaleoverheid.nl/kabinetsbeleid-digitalisering/werkagenda/>
- Emerson RM, Fretz RI, Shaw LL (2001) Participant observation and fieldnotes. In: Atkinson P, Coffey A, Delamont S, Lofland J, Lofland L (eds) Handbook of ethnography. SAGE Publications Ltd, pp 352–368
- Friedrich CJ (1963) Man and his government: an empirical theory of politics. McGraw-Hill
- Gemeente Den Haag (2020) Datagedreven werken voor onze stad. <https://denhaag.raadsinformatie.nl/document/8817239/1#search=%22datastrategie%202019%22>
- Gemeente Rotterdam (2023) Werken aan verantwoorde algoritmiseren: evaluatie algoritme-governance 2023. <https://gemeenteraad.rotterdam.nl/Agenda/Document/83d1e106-13ec-468d-9f84-26a862a1881a?documentId=e9990b23-e3c0-45b9-a623-9fd7e38512d&agendaItemId=d5e85430-f3de-4626-8609-a58737b2505c>
- Gemeente Utrecht (n.d.) Interne richtlijnen toepassing algoritmen. <https://utrecht.bestuurlijkeinformatie.nl/Document/View/166e5dae-e1db-4158-903b-71aa0e92b892>
- Goffey A (2008) Algorithm. In: Fuller M (ed) Software studies: a lexicon. MIT Press, pp 15–20

- Konaté S, Pali B (2023) “You have to talk with us, not about us”: exploring the harms of wrongful accusation on those affected in the case of the Dutch ‘childcare-benefit scandal’. *J Victimol* 16:139–164
- Medaglia R, Tangi L (2022) The adoption of artificial intelligence in the public sector in Europe: drivers, features, and impacts. In: *ACM International Conference Proceeding Series*, pp 10–18. <https://doi.org/10.1145/3560107.3560110>
- Meijer A (2005) Risk maps on the Internet: transparency and the management of risks. *Inform Polity* 10(1–2):105–113. <https://doi.org/10.3233/ip-2005-0062>
- Meijer A, Grimmelikhuijsen S (2020) Responsible and accountable algorithmization: how to generate citizen trust in governmental usage of algorithms. In: Schuilenburg M, Peeters R (eds) *The algorithmic society: technology, power, and knowledge*. Routledge, pp 53–66
- Meijer A, Lorenz L, Wessels M (2021) Algorithmization of bureaucratic organizations: using a practice lens to study how context shapes predictive policing systems. *Public Adm Rev* 81(5):837–846. <https://doi.org/10.1111/puar.13391>
- Mikalef P, Fjørtoft SO, Torvatn HY (2019). Artificial intelligence in the public sector: a study of challenges and opportunities for norwegian municipalities. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, pp 267–277. https://doi.org/10.1007/978-3-030-29374-1_22
- Mikalef P, Lemmer K, Schaefer C, Ylinen M, Fjørtoft SO, Torvatn HY, Gupta M, Niehaves B (2022) Enabling AI capabilities in government agencies: a study of determinants for European municipalities. *Gov Inform Q* 39(4). <https://doi.org/10.1016/j.giq.2021.101596>
- Ministerie van Binnenlandse Zaken en Koninkrijksrelaties (2021, June 10) Kamerbrief Voortgang AI en algoritmen. <https://www.rijksoverheid.nl/documenten/kamerstukken/2021/06/10/kamerbrief-voortgang-algoritmen-en-artificiele-intelligentie>
- Ministerie van Economische Zaken en Klimaat (2018, June 1) Nederlandse Digitaliseringsstrategie. <https://www.rijksoverheid.nl/documenten/kamerstukken/2021/06/10/kamerbrief-voortgang-algoritmen-en-artificiele-intelligentie>
- Ministerie van Economische Zaken en Klimaat (2019, October 8) Strategisch Actieplan voor Artificiële Intelligentie. <https://www.rijksoverheid.nl/documenten/beleidsnotas/2019/10/08/strategisch-actieplan-voor-artificiele-intelligentie>
- Ministerie van Justitie en Veiligheid (2021) Richtlijnen voor het toepassen van algoritmen door overheden en publieksvoorlichting over data-analyses. <https://www.rijksoverheid.nl/documenten/richtlijnen/2021/09/24/richtlijnen-voor-het-toepassen-van-algoritmen-door-overheden-en-publieksvoorlichting-over-data-analyses>
- Mol A (2002) *The body multiple: ontology in medical practice*. Duke University Press
- Moore MH (1995) *Creating public value: strategic management in government*. Harvard University Press
- Moore MH (2013) *Recognizing public value*. Harvard University Press
- Orlikowski WJ (2007) Sociomaterial practices: exploring technology at work. *Organ Stud* 28(9):1435–1448. <https://doi.org/10.1177/0170840607081138>
- Overheidsbreed Beleidsoverleg Digitale Overheid (2020) NL Digibeter 2020: Agenda Digitale Overheid. <https://www.digitaleoverheid.nl/wp-content/uploads/sites/8/2021/12/nl-digibeter-2020.pdf>
- Overman S, Schilleman T (2021) Toward a public administration theory of felt accountability. *Public Adm Rev* 82(1):12–22. <https://doi.org/10.1111/puar.13417>
- Raad voor het openbaar bestuur (2021, May 25) Sturen of gestuurd worden: advies over de legitimiteit van sturen met data. <https://www.raadopenbaarbestuur.nl/documenten/publicaties/2021/05/25/advies-sturen-of-gestuurd-worden>
- Rekenkamer Metropool Amsterdam (2023) Algoritmen: onderzoeksrapport. <https://publicaties.rekenkamer.amsterdam.nl/algoritmenonderzoeksrapport/>
- Rekenkamer Rotterdam (2021) Gekleurde technologie. <https://rekenkamer.rotterdam.nl/onderzoeken/algoritmes/>

- Seaver N (2017) Algorithms as culture: some tactics for the ethnography of algorithmic systems. *Big Data Soc* 4(2). <https://doi.org/10.1177/2053951717738104>
- van Eck M (2018) Geautomatiseerde ketenbesluiten & rechtsbescherming: Een onderzoek naar de praktijk van geautomatiseerde ketenbesluiten over een financieel belang in relatie tot rechtsbescherming. Doctoral Dissertation, University of Tilburg. https://pure.uvt.nl/ws/portalfiles/portal/20399771/Van_Eck_Geautomatiseerde_ketenbesluiten.pdf
- VNG (2019) VNG Realisatie Rapport. https://vng.nl/sites/default/files/2020-08/definitief_rapport_impactanalyse_who_v1.0.pdf
- Wetenschappelijke Raad voor het Regeringsbeleid (WRR) (2011) iOverheid. <https://www.wrr.nl/binaries/wrr/documenten/rapporten/2011/03/15/ioverheid/ioverheid.pdf>
- Wieringa MA (2020) What to account for when accounting for algorithms: a systematic literature review on algorithmic accountability. In: ACM Conference on Fairness, Accountability, and Transparency (FAT* '20), pp 1–18. <https://doi.org/10.1145/3351095.3372833>
- Wieringa MA (2022) “Hey SyRI, tell me about algorithmic accountability”: lessons from a landmark case. *Data Policy*, 5. <https://doi.org/10.1017/dap.2022.39>
- Williams M, Moser T (2019) The art of coding and thematic exploration in qualitative research. *Int Manage Rev* 15(1):45–55
- Zouridis S, Van Eck M, Bovens M (2020) Automated discretion. In: Evans T, Hupe P (eds) *Discretion and the quest for controlled freedom*. Palgrave Macmillan, pp 313–329
- Zuurmond A (1994) *De infocratie: Een theoretische en empirische heroriëntatie op Weber's ideaaltype in het informatietijdperk*. Phaedrus.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.



The Role of Technology in Citizens' Right to Good Administration: Examining the Impact of Smart Governments



Barbara Lazarotto

Abstract This chapter examines the implementation and impact of “smart governments” within the European Union, focusing on the integration of digital technology in public administration. It investigates the operational strategies adopted by these digital governments and their implications for the fundamental right to good administration. The study explores the increasing reliance on technology, driven by neoliberal policies, leading to the automation and digitalization of public services. While these advancements promise enhanced efficiency, transparency, and accessibility, they also pose significant risks to citizens’ rights, particularly concerning privacy, data protection, and meaningful citizen participation. This chapter critically analyzes the tension between the drive for efficiency through automation and the safeguarding of democratic values and fundamental rights. It advocates for a balanced approach that harmonizes technological innovation with the preservation of citizens’ rights and democratic principles.

1 Introduction

Public administrations from different countries have adopted a diverse range of technological applications to expedite and enhance the efficiency and effectiveness of public sector tasks. This integration of technology has had far-reaching effects, occasionally substituting analog tasks with digital counterparts, including those traditionally performed by public servants. Motivated by the imperative for digitalization, a substantial volume of data is continuously collected, processed, and utilized in various functions such as AI-driven automated decision-making processes and

Funding: The author has received funding from European Union’s Horizon 2020 research and innovation programme Legality Attentive Data Scientists under the GA n. 956562.

B. Lazarotto (✉)

Law, Science, Technology and Society (Lsts) Research Group, Vrije Universiteit Brussel, Brussels, Belgium

e-mail: Barbara.Da.Rosa.Lazarotto@vub.be

© The Author(s) 2025

J. Goossens et al. (eds.), *Public Governance and Emerging Technologies*,
https://doi.org/10.1007/978-3-031-84748-6_3

fraud detection. Governments often justify their pursuit of rationalization and optimization of tasks as a means to fulfill their duty to implement their citizens' right to a good administration. The digitization of functions can result in prompt government responses, ultimately leading to improved governance and enhanced services for citizens.

The transition toward ubiquitous automation in the public sector is here referred to as “smart governments,” a continuous process of digitalization adopted by governments, which has led to the automation of public activities, especially the ones in connection with citizens, such as the adoption of chatbots, automated fraud detection, and social benefit assessments. The evolution of this process has significantly altered the operations and obligations of public sector entities, potentially resulting in adverse effects on citizens. This includes a decline in face-to-face engagement with public officials, leading to a heightened feeling of alienation and decreased oversight in the administration of public services. Additionally, the increased reliance on automated processes has given rise to uncertainties regarding bureaucratic procedures, potentially causing confusion and frustration among citizens, especially those with low levels of digital literacy. Moreover, the shift toward automation has raised concerns about the potential erosion of citizens' rights, particularly in terms of privacy and data security, as well as the lack of transparency regarding the collection and utilization of citizens' data by governmental entities.

In light of these developments, this chapter aims to conduct an analysis of the “smart governments” phenomenon in the European Union context, scrutinizing its practices and their implications under the light of the right to a good administration as it is enshrined in the EU Charter of Fundamental Rights. To make this point, the remainder of the chapter will proceed as follows: First, Sect. 2 presents an overview of smart governments and their use of multiple technologies to enhance public efficiency based on neoliberal premises. Subsequently, Sect. 3 delves into the right to good administration, exploring how it can be used to enhance the protection of citizens in the context of the digitalization of governments. Section 4 continues the analysis by exploring the negative implications of smart governments, especially in diminishing the rights of citizens. Section 5 concludes the chapter with a final analysis.

2 Unpacking Smart Governments

Governments' reliance on information as a valuable resource for decision-making is not a new phenomenon. In the Roman Empire, record-keeping was essential for population counts, taxation, and representation of the population, animals, and soldiers. Records were also essential for the Empire's economy, being used to keep track of land, olives, crop fields, vineyards, and pastures. While the use of information has always been significant in the public sector, with technological developments, the use of data by governments has gained unprecedented weight. Now, we find ourselves in the era of a widespread digitalization of governments, a process

where tasks that were once handled in an analog manner are digitalized at a scale and speed never seen before. This process is defined as “smart governments.”

It should be noted that the digitalization process of the public sector did not occur in a single step. Instead, it was a non-linear development of the gradual implementation of multiple technologies. It started with the digitalization of documents and the introduction of computers as a form of support for public servants' work. At this stage, the use of technological tools was restricted as a means to aid public administration internally, introducing a routine in the decision-making process of public servants.¹ Later, with the popularization of the internet and population increase, governments adopted technology to enhance communication with citizens, developing websites and online communication systems.^{2,3}

Yet, in the late 1980s and early 1990s, there was a global shift towards neoliberal ideals that mainly deemed public administrations as outdated, slow, and inefficient. As this movement gained popularity, the public sector went through a series of reforms, mainly led by the pursuit of austerity policies that pushed for the reduction of public budgets while increasing efficiency and productivity.⁴ In the same period, the administrative theory of New Public Management (NPM) was introduced by the political scientist Christopher Hood.⁵ This theory emphasizes that the use of managerial practices—e.g., performance measures, indicators, targets, and audits—by public administrations makes it more effective and more cost-efficient. These beliefs catalyzed a series of reforms, infiltrating NPM's beliefs across public administrations.⁶ Since then, a process of gradual automation of tasks has been underway, replacing professional workers with computer-based decision systems.⁷ This development has required a mass collection and processing of data through the use of a network of connected physical objects with sensors, actuators, and other technologies to convert physical phenomena into a digital signal, later collected and transmitted in real-time in multiple areas such as transport, energy, urban solutions, and defense.⁸ This data is later processed with the use of multiple technologies such as machine learning systems and predictive analytics, as well as natural language processing, amongst others. Using these multiple technologies, the public sector can not only understand societal patterns, trends, and policy impacts but also help in day-to-day operational decisions and support public policies.⁹

Against this backdrop, the conjunction of technological advancements and socio-economic factors fostered fertile soil that led to the development of what we call

¹ See Inbar (1979).

² See Ranchordás and Scarcella (2021).

³ See Lee-Archer (2023).

⁴ Dencik (2022), pp. 145–165.

⁵ See Hood (1995).

⁶ Harlow (1997), pp. 245–261.

⁷ See Bovens and Zouridis (2002).

⁸ See Barcevičius et al. (2019).

⁹ See Bachner and Hill (2014).

“smart governments.”¹⁰ Smart governments can be defined as a type of governance that uses multiple technologies to implement digital automation, algorithm and AI-driven decision-making, and data-driven technologies directed toward optimizing public administrations.¹¹ In this context, technologies can support public servants’ decision-making, perform risk detection and assessment tasks through scoring and sorting, and optimize and coordinate funds and human resources through the automatic collection and processing of data both from citizens and from the environment.¹² Smart governments can also be seen as an evolution of New Public Management with a foundation in technology acting as a catalyst for change with an underlying neoliberal political view, which differentiates smart governments from other terms that aim to refer to the use of technology by governments with no underlying political connotation, such as e-government.¹³

One of the main claims for the widespread digitalization of public administration is the use of predictive analytics, which uses a combination of data mining, statistics, and modeling to make predictions and trace expected and unexpected behaviors. Predictive analytics can also be used to better shape public policies, optimize the deployment of resources, make the delivery of services more efficient, and prepare possible interventions before problems occur.¹⁴ Predictive analytics are mostly used in governments in areas such as public safety,¹⁵ education, public health,¹⁶ and fraud detection.¹⁷ In addition, these technologies would allow the full or partial automation of public servant’s tasks and functions of public servants with automation, reducing work overflow, decreasing errors and inconsistencies, and improving the speed and accuracy of service delivery.¹⁸ By replacing human servants in certain tasks, they would be freed to more complex tasks, reducing human resources expenses.

These digital tools, to retain their dynamic utility, also require an interconnected and continuous flow of information in the public sector, resulting in the process of dataset integration and data sharing between multiple governmental agencies in a process described by Bovens and Zouridis as “system-level bureaucracy.”¹⁹ A key feature of this system integration is the connection of once-siloed datasets into a single point of access to electronic services and information in the public sector. This facilitates not only cross-departmental informational flows but also the delivery of public services to citizens, creating “one-stop shops,” which can be described

¹⁰ See Yeung (2023).

¹¹ See Janowski (2015); Veale and Brass (2019).

¹² See Yeung (2023).

¹³ Baptista (2005), p. 170.

¹⁴ See Bright et al. (2019).

¹⁵ See Tomar et al. (2016).

¹⁶ Qureshi (2014).

¹⁷ See De Fremery (2018).

¹⁸ See Barcevičius et al. (2019).

¹⁹ See Bovens and Zouridis (2002).

as a single digital counter where citizens can connect with the public sector to solve any type of matter.²⁰ Thus, the work of public servants shifts from handling individual cases to the optimization of processes through system development.²¹ While some results of technological adoption may be proven true, it is also necessary to emphasize that human resources expenses may also increase due to the demand to increase digital literacy amongst the officials that remain. Whereas once public servants could go by with their daily tasks by only learning basic computing skills, now another level of literacy is required. So-called “second literacies,” “new literacies,” or “contemporary literacies” demand not only knowing how to use basic technologies but also include a set of skills that allow servants to recognize when they need information, find it, interpret it, and exploit it correctly. Thus, this level of new literacy will undoubtedly demand further human resources and qualifications from the public sector.²²

In Germany's *Bundesagentur für Arbeit*—BA (Germany's Federal Labor Agency) is said to have reduced the number of fraudulent claims by 20% with the use of data science techniques and predictive models.²³ In addition, these technologies would allow the full or partial automation of public servant's tasks and functions of public servants with automation, reducing work overflow, decreasing errors and inconsistencies, and improving the speed and accuracy of service delivery.²⁴ By replacing human servants in certain tasks, they would be freed to more complex tasks, reducing human resources expenses. While some results of technological adoption may be proven true, it is also necessary to emphasize that human resources expenses may also increase due to the demand to increase digital literacy amongst the officials that remain. Whereas once public servants could go by with their daily tasks by only learning basic computing skills, now another level of literacy is required. So-called “second literacies,” “new literacies,” or “contemporary literacies” demand not only knowing how to use basic technologies but also include a set of skills that allow servants to recognize when they need information, find it, interpret it, and exploit it correctly. Thus, this level of new literacy will undoubtedly demand further human resources and qualifications from the public sector.²⁵

In this context, depending on how technology is adopted and embedded in work practices, technology can dangerously infiltrate public administration. Management systems and digital decision trees may place strict rules and boundaries on officials' work, putting the ultimate decision in the hands of the ones who design and deploy the software instead of in the hands of public servants.²⁶ Moreover, by over-relying on a digital assemblage as a form of social order and mode of governance, smart

²⁰ See Ranchordás (2022).

²¹ See Cleary and Tax (2011).

²² See Ranchordás (2022).

²³ See Cleary and Tax (2011).

²⁴ See Barcevičius et al. (2019).

²⁵ See American Library Association (1989); Herout (2015).

²⁶ See Bovens and Zouridis (2002).

governments reshape public values, placing internal organizational values such as performance and efficiency as priorities concerning traditional public sector values such as transparency and the protection of fundamental rights.²⁷

This movement undoubtedly changes the nature of the relationship between government and citizens, replacing one-on-one relationships and meetings with cameras, sensors, websites, and chatbots. This change has a direct impact on citizens at multiple levels, especially those directly affected by government actions. This topic will be further explored in the following section.

3 Right to Good Administration

The relationship between governments and citizens has been permeated by discussions related to the limits of governmental power and the rights of citizens. Multiple social contract theories were developed to describe the origins and legitimacy of political authority, such as the theories of John Locke and Rousseau, who described the formation of governments as a surrender of some individual rights in exchange for an orderly and secure society.²⁸ In the context of social contract theories, it is the duty of governments to seek the public good while nurturing an environment where citizens can enjoy equality and freedom under the law.²⁹ It was under this framework that classic eighteenth-century beliefs of legality, equality, and certainty were transported to the public sector in the welfare state, developing what would become procedural and material principles of good administration, keeping them within the bounds of the rule of law.³⁰ Therefore, administrative rules play an essential role in protecting the fundamental rights of citizens. It is in this framework that the right to good administration has emerged.

Drawing from legal traditions from multiple European Member States, such as Spain and Portugal, which included specific provisions to protect individuals from unfair administrative procedures, the EU went through a reform that included principles of administrative law in the EU Charter of Fundamental Rights.³¹ The right to a good administration is expressed in Article 41 of the EU Charter of Fundamental Rights as a subjective individual fundamental right that not only requires administrators to act in accordance with the law but also provides citizens with an enforceable right.³² Under the scope of the right to a good administration, Article 41 mentions that every person has the right to have his or her affairs handled impartially, fairly, and within a reasonable time. The second paragraph of the article

²⁷ See Hickok (2022).

²⁸ See Riley (1982).

²⁹ See Garth-James (2021).

³⁰ See Bovens and Zouridis (2002).

³¹ See Dutheil de la Rochère (2008).

³² See Kristjánsdóttir (2013).

includes the right to be heard before any individual measure is taken, the right of every person to access his or her file, and the obligation for administrative bodies to give reasons for their decisions.

It is especially important to note that the European Court of Justice had an important role in the development of the right to good administration. This trajectory started in the European Court of Justice landmark case in *Technische Universität München v Hauptzollamt München-Mitte*³³, in which the Court emphasized the importance of respecting procedural guarantees in administrative procedures and that non-compliance with these guarantees leads to the illegality of the decision.³⁴ Thus, the right to a good administration under Article 41 presents itself as an umbrella right that is not limited to the rights enumerated in it, such as the duty of the administrative sector to reply to requests, collect sufficient information, and examine requests in a fair manner, but also encompassing both procedural and substantive requirements.³⁵

However, while the Charter's provisions are applicable to the institutions and bodies of the Union and the Member States when implementing Union law, Article 41 has omitted its applicability to the Member States, having a narrower application than the rest of the Charter, only being applicable to the institutions and bodies of the European Union.³⁶ At the same time, since the Treaty of Lisbon, which was entered into force in December 2009, the Charter has been a legally binding document, the same as a Treaty. Hence, there is an incompatibility between the limited scope of Article 41 and the Charter's objective, which is to strengthen the protection of fundamental rights.³⁷ Article 41 seems to be compatible with the doctrine of procedural autonomy, which grants member states the freedom to establish their own national procedural rules to govern the exercise of EU law within their territories.³⁸ Under this theory, Member States do not have to take the principle of good administration into account when acting within their administrative obligations. However, the theory of procedural autonomy does not apply to the application of general principles of EU law. In this case, Member States must reflect the general principles as minimum standards.³⁹ This, from our point of view, is the case with the principle of good administration since Article 253 of the Treaty of Lisbon mentions that "legal acts shall state the reasons on which they are based and shall refer to any proposals, initiatives, recommendations, requests or opinions required by the Treaties." Therefore, despite the narrow scope of Article 41 of the EU Charter of Fundamental Rights, it

³³ Case C-269/90, *Hauptzollamt München-Mitte v. Technische Universität München* [1991]. <https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:61990CJ0269>.

³⁴ See Nehl (2009).

³⁵ See Pech and Groussot (2010).

³⁶ See Kristjánsson (2013).

³⁷ See Groussot et al. (2011).

³⁸ See Ortlep and Verhoeven (2012).

³⁹ See Reichel (2008).

does not interfere with the applicability of the general EU principle of good administration to Member States as a fundamental principle to be protected.

In the context of smart governments, while governments often justify the extensive use of technology as a means to fulfill the principle of good administration through the delivery of an efficient, seamless public service, they may result in harmful consequences such as biased and opaque outcomes from automated decision-making that do not comply with the level of accountability and transparency required from the public sector. Against this informational background, it appears that the principle of good administration enshrined in the CFR is not just a justification for the digitalization of public administration but an umbrella principle in the protection of citizens' rights. Under the scope of the principle of good administration, cross-cutting legally enforceable and unenforceable guarantees must be developed to counterbalance the digitalization of public sector decision-making systems with personal dignity, autonomy, and freedom as pillars. To begin with, the actions of administrative bodies must not be founded in a rigid *dura lex sed lex* legality of automated decision-making in a binary fashion. Instead, good administration asks for a public administration that is transparent, open, and democratic in its decision-making.⁴⁰ In this regard, the recognition of fundamental values such as dignity, autonomy, and freedom must be done by developing individual procedural safeguards that are adapted to the digital 21st-century information society.⁴¹ However, this recognition process goes beyond the direct relationship between citizens and governments; in this context, we propose a multi-level approach to the principle of good administration, which encompasses multiple levels of government, including the design, development, maintenance, selection, procurement, and use of AI systems.⁴²

The first layer of implementation of the principle of good administration should be focused on the design of the technological systems, their deployment, and maintenance in the public sector.⁴³ This infiltration process must encompass the process of technological procurement and design that must be adequate to the public sector's duties of diligence.⁴⁴ The development and maintenance of technological systems used in the public sector must involve different stakeholders, including citizens, academia, and civil society, as a form of making the adoption process more democratic. Moreover, the public sector must be transparent about *if* and *how* technology is *needed* and adopted throughout the public administration and how information

⁴⁰An example of good practice is the one implemented by the Social Insurance Bank (SVB) in The Netherlands named *Garage de Bedoeling*, translated in English as 'Garage of Intention'. The Garage has the main objective of discussing and handling cases that do not fit in a rigid legality, drawing the law's intention as opposed to its strict wording. See <https://magazines.rijksoverheid.nl/adr/adrover/2019/03/garage-de-bedoeling>.

⁴¹Citron (2007), p. 1249.

⁴²This multi-level process includes algorithmic accountability See Wieringa (2020), pp. 1–18 and Shah (2018).

⁴³See Veale and Brass (2019).

⁴⁴See Hickok (2022).

systems are linked, allowing individuals to comprehend how their personal data is processed within the system and who can have access to it, including if it is shared with other governments or outside actors.

The second layer of implementation of the principle of good administration is the enactment of crosscutting rights that are aimed at enhancing transparency in the technological decision-making process. First, public servants must be trained to enhance their capacity to comprehend how the technologies they use in the public sector work, including technical aspects of datasets and possible harmful outcomes.⁴⁵ Furthermore, the ground and considerations of public administration decisions must be fully disclosed, giving a satisfactory answer to citizens that is not merely grounded in a binary decision of a machine but balanced between individual and administrative costs.⁴⁶ Equally important is to provide citizens with tools that allow them to challenge administrative decisions and be in an equal position in relation to governments. the “human-in-the-loop” might be one of the possible measures to insert more nuance into automated decision-making, it does not “cleanse” away problematic decisions, especially considering the tendency of public servants to over-relying on computer decisions.⁴⁷ Therefore, the right to be heard must be included in any procedure that might culminate in a decision of an administrative or judicial nature that affects a person’s interest, including automated decision-making.⁴⁸

Finally, this multi-layered approach that has the principle of good administration as a background must be translated into codification to comply with the Continental positivist tradition of developing legal rules. Thus, by transforming the principle of good administration into an *enforceable right* to good administration within the legal landscapes of Member States, these principles will be a legally enforceable objective guarantee upon which individuals can rely to protect their rights. The codification of a right to a good administration would be in tune with the necessity of checks and balances tools to protect citizens against potential governmental abuses, which may come disguised as oversimplification and digitalization of administrative practices. The Artificial Intelligence Act, which was recently approved in the EU at the time of this study, may be considered a first step toward this direction, yet it has some caveats. The AI Act takes a cross-sectoral approach by tailoring levels of risk from unacceptable to minimal.⁴⁹ However, as many have stressed, the Act has limited rights and redress mechanisms, not contemplating or providing mechanisms for individuals or groups to fight against the negative impacts of AI systems.⁵⁰

⁴⁵ See Veale and Brass (2019).

⁴⁶ See Bovens and Zouridis (2002).

⁴⁷ Crootoft et al. (2023), p. 482.

⁴⁸ First Chamber of CJEU, Case C-277/11, Judgement 22 November 2012.

⁴⁹ See Prainsack and Forgó (2024).

⁵⁰ See Ogunleye (2022).

4 Navigating the Negative Implications of Smart Governments

Despite the multiple advantages promised by smart governments, such as the reduction of bottlenecks of routine administrative tasks, it is important to note that they also have multiple negative outcomes. They can directly affect not only the foundations of the public sector and public administration but also individuals who are subject to government actions that are, whether directly or indirectly, encouraged by the reliance of government departments on digitized practices. Hence, this analysis aims to examine not only the internal implications but also those that extend beyond the public sector, such as its impacts on the lives of citizens, towards a comprehensive understanding of its negative effects on society.

The dynamics of the relationship between governments and their citizens are inherently asymmetrical, a topic thoroughly examined by various social contract theories throughout the centuries as explored by Thomas Hobbes, John Locke, Rousseau, Kant, and Hegel.⁵¹ The fulcrum of this unbalanced relationship lies in the inherent authority vested in the government, which is granted by the citizens themselves, who willingly surrender certain freedoms and authority to the government in exchange for protection, order, and the assurance of basic rights through a social contract. These social contract theories eventually created a fertile ground for the development of a system of checks and balances, as initially crafted by Rousseau, designed to shield citizens from potential abuses of power by the government by setting clear boundaries on the extent of governmental power.⁵² The preservation by the government of these rights and principles represents the commitment by the government to operate within defined parameters and a commitment by citizens to contribute to a functioning society.

Therefore, a delicate balance must be found between different public values, namely efficiency, effectiveness, transparency, and due process. However, when a government places digitization and efficiency efforts above traditional and key public sector values, there is a pressing need to oversimplify (and often distort) processes and procedures, so they fit into code parameters that demand simplified formats such as simple questions and answers.

This leads to various adverse consequences. The first observable one is the loss of checks and balances procedures—such as division of power, oversight mechanisms, and due process instruments—within the public administration due to the need to simplify procedures and automate tasks. This results in the loss of context information that cannot be readily datafied,⁵³ essential aspects to restrain potential governmental abuse of power and protect citizens from these abuses.⁵⁴ This was

⁵¹ See Riley (1982).

⁵² See Sasan (2021); Holcombe (2018).

⁵³ On the importance of context and the difficulty of “datafying” certain aspects of human life, see Kitchin (2021) and Drucker (2011).

⁵⁴ See Holcombe (2018).

visible in the infamous case of *Systeem Risico Indicatie* (SyRI) in The Netherlands, a system that heavily relied on data from different administrative organs to detect fraud in social welfare benefits. The case gained notoriety for using data to detect discrepancies (such as double citizenship and owning a car) to label people as potential fraudsters, violating their human rights.⁵⁵

As Citron has highlighted, the preference for a simplification policy over nuanced analysis by itself narrows the field for the expertise model of administrative law;⁵⁶ this tendency eliminates the grey zones imposed by the complexity of “real life,” such as the double nationality example, leading to an unforgiving government that reduces administrative tasks to a cold objectivity. Here, arbitrariness is addressed by excessive rigidity, leaving no room for *Eenzelfallgerechtigheit*, translated as justice in each particular case.⁵⁷ This is especially problematic when it comes to vulnerable individuals and communities, who, due to their conditions, are more affected by governmental actions and demand a more nuanced treatment by the public sector, in the case of SyRI, women and immigrant families.

This demonstrates how the delegation of core public tasks to automation has the deep potential to transform how public authority is exercised, distributed, and experienced. While it is true that humans have their own preconceptions, and the introduction of automation into the decision-making process may remove discretion and arbitrary exercise of power,⁵⁸ it is important to note that automated systems may also result in biased results in the case of biased datasets they are trained on and how they are programmed. The mass collection of data also includes data that is “noisy” and inaccurate, with irregular and heterogeneous properties.⁵⁹ Not to mention the digital exclusion that is reflected in the underrepresentation of certain groups of individuals in datasets, exacerbating possible vulnerabilities with false positives.⁶⁰ This reliance, without critically examining the possible data quality issues and biases, exacerbates the spread of problematic results, creating a self-perpetuating cycle of inaccuracies and reinforcing existing prejudices.⁶¹ At the same time, automated decision-makers lack the inherent human factor that public servants bring to their roles. Despite the amount of data used to train models, the absence of personal experience and compassion limits automated systems' capacity to address often complex situations.⁶²

The lack of human decision-makers is especially problematic when it comes to vulnerable individuals who are treated as data points whose actions should be

⁵⁵ See Wieringa (2023); van Bekkum and Borgesius (2021).

⁵⁶ See Citron (2007).

⁵⁷ See Bovens and Zouridis (2002).

⁵⁸ Ibid.

⁵⁹ See Schintler and Kulkarni (2014).

⁶⁰ See Park and Humphry (2019).

⁶¹ See Citron (2007).

⁶² See Barth and Arnold (1999).

interpreted with the objective of optimizing public sector operations.⁶³ The United Nations Special Rapporteur on Extreme Poverty and Human Rights has criticized⁶⁴ the reforms in the British welfare system, which has a techno-solutionism tendency to go “digital by default,” removing any type of in-person appointment, as several vulnerable individuals who are claimants of social support do not have the required digital literacy or cannot afford the expenses of internet access at home.⁶⁵ The same was evident in a 2012 study conducted by the European Commission that demonstrated the lack of ability and trust as the main reasons why European citizens do not engage with governments through eGovernment channels.⁶⁶ Hence, by focusing on digitalization, governments ignore social and political aspects of society. The SyRI case, for instance, had a nefarious practice that was labeled as “institutional racism”.⁶⁷ As a result, families were pushed into poverty, and more than a thousand children were placed into foster care; some cases even pushed victims to suicide.⁶⁸

Additionally, the lack of transparency in automated systems is problematic since it shields them from the necessary scrutiny demanded by the public sector duties of accountability and transparency.⁶⁹ While the public sector can provide general information about automated systems, this transparency is often not sufficient for a deeper analysis of potential errors in a system, especially in the context of smart governments, which are complex ecosystems in which AI is one of the multiple technologies.⁷⁰ In the case of SyRI, the Hague District Court noted that the system lacked transparency measures since it did not allow individuals to track their personal data in the system. Instead, the system was opaque, intensifying risk.⁷¹ These opaque characteristics of automated systems become problematic when they are adopted in the context of public administration processes since they leave out essential tenets of justice such as the right to be informed, heard, and treated fairly in any administrative proceedings.

Equally important is the potential for security risks that come along with a wide collection of data, which may include sensitive information that encompasses all types of personal data from healthcare to financial information that can be used by the public sector to trace a complete picture of an individual’s life, not just undermining citizen’s privacy, but also furthering the power and knowledge imbalance

⁶³ See Yeung (2023).

⁶⁴ See United Nations (2019).

⁶⁵ See Ranchordas (2022).

⁶⁶ PUBLIC SERVICES ONLINE, ‘Digital by Default or by Detour?’ Assessing User Centric eGovernment performance in Europe – eGovernment Benchmark 2012” (DG Connect), available at <https://op.europa.eu/en/publication-detail/publication/23d5089b-4910-4cc5-b950-aecca2ffc4f8e>.

⁶⁷ See Davidson et al. (2022).

⁶⁸ See Heikkilä (2022).

⁶⁹ See Citron (2007).

⁷⁰ See Veale and Brass (2019).

⁷¹ NJCM et al. v The Dutch State (2020) The Hague District Court ECLI: NL: RBDHA:2020:1878 (SyRI). Available at: [https://uitspraken.rechtspraak.nl/inziendocument?id=4ECLI:NL:RBDHA:2020:1878\[6.90\]](https://uitspraken.rechtspraak.nl/inziendocument?id=4ECLI:NL:RBDHA:2020:1878[6.90]); van Bekkum and Borgesius (2021).

between the citizen and the government's power. The power imbalances caused by this "data deluge", as defined as Pencheva et al., gives governments enough power to monitor citizens, in an Orwellian scenario, besides opening the door to potential abuses and misuses of data for different purposes than initially intended. The Chinese scenario of governmental mass surveillance⁷² is an example of these practices, which merge censoring techniques, with data analytics for political surveillance purposes.⁷³

Overall, as it was possible to observe, the digitalization of governments has a significant impact on the relationship between governments and citizens. While it may enhance efficiency, accessibility and service delivery, it also can lead to the erosion of citizen's rights which act as barriers to possible governmental abuses such as surveillance and potential misuse of sensitive information. Therefore, there is a need for the development of a robust legal framework grounded in the right to a good administration which emphasizes multi layered rights that enhance the quality of public services, foster transparency, and strengthen democratic governance.

5 Concluding Thoughts and the Path Forward

The previous sections of this discourse have delved into the proliferation of technology in governmental operations, grounded in neoliberal ideals emphasizing efficiency and optimization in the public sector. This widespread use of technology in governments has led to the notion of smart governments, mainly characterized by multiple interconnected technologies to reach an efficient, optimized public sector. Yet, while promising enhanced governance, the widespread use of technology also brings forth a series of adverse consequences for the lives of citizens.

In response to these challenges, this chapter proposes a comprehensive, multi-layered system of protection within the public sector, anchored in the fundamental tenets of good administration. This strategic framework aims to empower citizens with an array of tools to counterbalance the overreaching tendencies of smart governments, thereby fortifying essential rights such as transparency and the right to be heard, such as the inclusion of citizen access to decision trees by the public sector and the augmentation of transparency measures in the procurement of digital technologies.

The recently approved AI Act might be a part of this multi-layered solution. The AI Act is a binding risk-based regulation that will apply across domains, including the public sector, which creates different levels of AI risk that range from unacceptable to minimal. Amongst the high-risk AI Systems, which are systems that require an additional list of obligations, is the use of AI in biometrics, critical infrastructure, education, employment, essential services, law enforcement, migration, and justice.

⁷² See Knockel et al. (2020).

⁷³ See Pencheva et al. (2020).

Amongst the additional requirements of high-risk AI systems is the establishment of a regularly updated and reviewed risk management system that is continuous throughout the AI lifecycle. Moreover, the system should pay considerable attention to whether the AI system could impact people under 18 or other vulnerable groups, as the SyRi case did. Lastly, the Fundamental Rights Impact Assessment for High-Risk AI Systems might play an interesting role in the case of smart government applications since it demands that public services assess how the AI system could impact people's fundamental rights.⁷⁴

While it is crucial to acknowledge that, given the constraints of a chapter-length discussion, the depth of this topic remains incompletely explored within this study, hopefully that this chapter serves as a first step toward demonstrating the imperative need for a set of governance structures and guarantees that align with the configurations of our digitized society.

References

- American Library Association (1989) Presidential Committee on Information Literacy: Final Report
- Bachner J, Hill K (2014) Advances in public opinion and policy attitudes research. *Policy Stud J* 42(S1):S51–S70. <https://doi.org/10.1111/psj.12052>
- Baptista M (2005) e-Government and state reform: policy dilemmas for Europe. *Electr J eGovernment* 3(4):167–174
- Barcevičius E, Cibaitė G, Codagnone C, Gineikytė V, Klimavičiūtė L, Liva G, Matulevič L, Misuraca G, Vanini I (2019) Exploring digital government transformation in the EU: analysis of the state of the art and review of literature. Publications Office of the European Union
- Barth TJ, Arnold E (1999) Artificial intelligence and administrative discretion: implications for public administration. *Am Rev Public Adm* 29(4):332–351. <https://doi.org/10.1177/02750749922064463>
- Bovens M, Zouridis S (2002) From street-level to system-level bureaucracies: how information and communication technology is transforming administrative discretion and constitutional control. *Public Adm Rev* 62(2):174–184. <https://doi.org/10.1111/0033-3352.00168>
- Bright J, Ganesh B, Seidelin C, Vogl TM (2019) Data science for local government. Oxford Internet Institute, Oxford
- Citron DK (2007) Technological due process. *Wash Univ Law Rev* 85(6):1249–1313. https://heinonline.org/HOL/Page?handle=hein.journals/walq85&div=38&g_sent=1&casa_token=r0de-bNvRKUIAAAAA:8lu5C4RTwHEDmIW4I0vt4jfS3AE8344GZm4gKKzhDK18IAkMyPYtKFTlqF3v4TKwyqsKcgcJ5g
- Cleary D, Tax RI (2011) Predictive analytics in the public sector: Using data mining to assist better target selection for audit. In: Proceedings of the European Conference on e-Government, Vol. 9(2). ECEG, pp 168–176
- Crotoft R, Kaminski ME, Price W, Nicholson II (2023) Humans in the loop. *Vanderbilt Law Rev* 76(2):429–510. https://heinonline.org/HOL/Page?handle=hein.journals/vanlr76&div=14&g_sent=1&casa_token=p2pEagOfyZEAAAAA:EPHvfDfV7GHFr_rF0Eon-aT9sg-k5FErcfUNUdln8Vw8Fif6w0-O7Ru96Md8wyi6lWj0QUIQUw

⁷⁴ See Prainsack and Forgo (2024).

- Davidson D, Geiger G, Schot E, Hijink M, Adriaens S, Bulman M, Konijn J, van der Woude A, Hekman L, Howden D (2022, December 20) The Algorithm Addiction. Mass profiling system SyRI resurfaces in the Netherlands despite ban and landmark court ruling. Lighthouse Reports. <https://www.lighthousereports.nl/investigation/the-algorithm-addiction/>
- Dencik L (2022) The datafied welfare state: a perspective from the UK. In: Hepp A, Jarke J, Kramp L (eds) *New perspectives in critical data studies. Transforming communications – studies in cross-media research*. Palgrave Macmillan, pp 145–165
- Drucker J (2011) Humanities approaches to graphical display. *Dig Human Quart* 5(1):1–21
- Dutheil de la Rochère J (2008) The EU Charter of Fundamental Rights, not binding but influential: the example of good administration. In: Arnall A, Eeckhout P, Tridimas T (eds) *Continuity and change in EU law. Essays in Honour of Sir Francis Jacobs*. Oxford University Press, pp 157–172
- Fremery R (2018, February 28) Big data and government: how the public sector leverages data insights. Hortonworks. Retrieved from <https://blog.cloudera.com/data-360/big-data-and-government-how-the-public-sector-leverages-data-insights/>
- Garth-James K (2021, January 19) The implications of social contract theory for public governance. *Public Administration Times*, American Society for Public Administration. <https://patimes.org/the-implications-of-social-contract-theory-for-public-governance/>
- Groussot X, Pech L, Petursson GT (2011) The scope of application of fundamental rights on Member States' action: in search of certainty in EU adjudication. Available at SSRN https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1936473
- Harlow C (1997) Back to basics: reinventing administrative law. *Public Law* 8(2):245–261
- Heikkilä M (2022, March 30) AI: decoded: a Dutch algorithm scandal serves a warning to Europe – the AI Act won't save us. *Politico*. <https://www.politico.eu/newsletter/ai-decoded/a-dutch-algorithm-scandal-serves-a-warning-to-europe-the-ai-act-wont-save-us-2/>
- Herout L (2015) Computer literacy in public administration. In: *ICERI2015 Proceedings*. IATED, pp 5965–5973
- Hickok M (2022) Public procurement of artificial intelligence systems: new risks and future proofing. *AI Soc* 39:1213–1227. <https://doi.org/10.1007/s00146-022-01572-2>
- Holcombe RG (2018) Checks and balances: enforcing constitutional constraints. *Economies* 6(4):57–69. <https://doi.org/10.3390/economies6040057>
- Hood C (1995) The “new public management” in the 1980s: variations on a theme. *Account Org Soc* 20(2-3):93–109. [https://doi.org/10.1016/0361-3682\(93\)E0001-W](https://doi.org/10.1016/0361-3682(93)E0001-W). <https://www.ala.org/acrl/publications/whitepapers/presidential>
- Inbar M (1979) *Routine decision-making: the future of bureaucracy*. SAGE Publications
- Janowski T (2015) Digital government evolution: from transformation to contextualization. *Gov Inform Quart* 32(3):221–236. <https://doi.org/10.1016/j.giq.2015.07.001>
- Kitchin R (2021) *The data revolution: a critical analysis of big data, open data and data infrastructures*. SAGE Publications
- Knockel J, Parsons C, Ruan L, Xiong R, Crandall J, Deibert R (2020) We chat, they watch: how international users unwittingly build up WeChat's Chinese censorship apparatus. *The Citizen Lab*
- Kristjánssdóttir MV (2013) Good administration as a fundamental right. *Icelandic Rev Polit Adm: Stjórnsmál og stjórnsýsla*, 9(1):237–255. <https://doi.org/10.13177/irpa.a.2013.9.1.12>
- Lee-Archer B (2023) Effects of digitalization on the human centricity of social security administration and services. *ILO Working Paper*. <https://doi.org/10.54394/PMPD3825>
- Nehl HP (2009) Good administration as procedural right and/or general principle? In: Hofmann HCH, Türk AH (eds) *Legal challenges in EU administrative law: towards an integrated administration*. Edward Elgar Publishing, pp 322–351
- NJCM et al. v. The Dutch State (2020) The Hague District Court ECLI: NL: RBDHA:2020:1878 (SyRI). <https://uitspraken.rechtspraak.nl/inziendocument?id=ECLI:NL:RBDHA:2020:1878>
- Ogunleye I (2022) AI's redress problem: recommendations to improve consumer protection. *CLTC White Paper Series*. CLTC Berkeley

- Ortlep R, Verhoeven MJM (2012) The principle of primacy versus the principle of national procedural autonomy. *Netherlands Adm Law Libr* 2012:1–16. <https://doi.org/10.5553/NALL.000004>
- Park S, Humphry J (2019) Exclusion by design: intersections of social, digital and data exclusion. *Inform Commun Soc* 22(7):934–953. <https://doi.org/10.1080/1369118X.2019.1606266>
- Pech L, Groussot X (2010) Fundamental rights protection in the EU post Lisbon treaty. Available at SSRN https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1628552
- Pencheva I, Esteve M, Mikhaylov SJ (2020) Big data and AI—A transformational shift for government: so, what next for research? *Public Policy Adm* 35(1):24–44. <https://doi.org/10.1177/0952076718780537>
- Prainsack B, Forgó N (2024) New AI regulation in the EU seeks to reduce risk without assessing public benefit. *Nat Med* 30(5):1235–1237. <https://doi.org/10.1038/s41591-024-02874-2>
- Qureshi B (2014) Towards a digital ecosystem for predictive healthcare analytics. In: *Proceedings of the 6th International Conference on Management of Emergent Digital EcoSystems*. ACM, pp 34–41
- Ranchordás S (2022) The digitization of government and digital exclusion: setting the scene. In: Blanco de Moraes C, Ferreira Mendes G, Vesting T (eds) *The rule of law in cyberspace*. Springer International Publishing, pp 125–148
- Ranchordás S, Scarella L (2021) Automated government for vulnerable citizens: intermediating rights. *William & Mary Bill Rights J* 30(2):373
- Reichel J (2008) Between supremacy and autonomy - applying the principle of good administration in the member states. In: Bernitz U, Nergelius J, Cardner C (eds) *General principles of EC law in a process of development: reports from a conference in Stockholm, 23–24 March 2007, Organised by the Swedish Network for European Legal Studies*. Wolters Kluwer Law & Business
- Riley P (1982) *Will and political legitimacy: a critical exposition of social contract theory in Hobbes, Locke, Rousseau, Kant, and Hegel*. Harvard University Press
- Sasan JMV (2021) The social contract theories of Thomas Hobbes and John Locke: comparative analysis. *Shanlax Int J Arts Sci Human* 9(1):34–45. <https://doi.org/10.34293/sijash.v9i1.4042>
- Schintler LA, Kulkarni R (2014) Big data for policy analysis: the good, the bad, and the ugly. *Rev Policy Res* 31(4):343–348. <https://doi.org/10.1111/ropr.12079>
- Shah H (2018) Algorithmic accountability. *Philos Transact Royal Soc Math Phys Eng Sci* 376(2128):1–6. <https://doi.org/10.1098/rsta.2017.0362>
- Tomar L, Guicheney W, Kyarisiima H, Zimani T (2016) *Big data in the public sector: selected applications and lessons learned*. Inter-American Development Bank
- United Nations (2019 April 23) *Report of the Special Rapporteur on extreme poverty and human rights: visit to the United Kingdom of Great Britain and Northern Ireland*. United Nations, Available at <https://undocs.org/en/A/HRC/41/39/Add.1>
- van Bekkum M, Borgesius FZ (2021) Digital welfare fraud detection and the Dutch SyRI judgment. *Eur J Soc Secur* 23(4):323–340. <https://doi.org/10.1177/13882627211031257>
- Veale M, Brass I (2019) Administration by Algorithm? Public management meets public sector machine learning. In: Yeung K, Lodge M (eds) *Algorithmic regulation*. Oxford University Press, pp 121–149
- Wieringa M (2020) What to account for when accounting for algorithms: a systematic literature review on algorithmic accountability. In: *Proceedings of the 2020 conference on fairness, accountability, and transparency*. Association for Computing Machinery, pp 1–18
- Wieringa M (2023) “Hey SyRI, tell me about algorithmic accountability”: lessons from a landmark case. *Data Policy* 5:e2-1–e2-24. <https://doi.org/10.1017/dap.2022.39>
- Yeung K (2023) The new public analytics as an emerging paradigm in public sector administration. *Tilburg Law Rev* 27(2):1–32. <https://doi.org/10.5334/tlrl.303>

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.



Explainable AI Is No Silver Bullet: Towards a Contextual Understanding of Appropriate Reliance on AI in Law Enforcement



Merijn Bruijnes, Stephan Grimmelikhuijsen, and Marcel Robeer

Abstract Police officers increasingly rely on AI recommendations in their decision-making processes. This chapter delves into the potential and challenges of AI in law enforcement, focusing on the role of Explainable AI (XAI) on appropriate reliance, where appropriate is neither stubbornly rejecting, nor complacently accepting AI recommendations. We discuss recent XAI research findings and their implications for law enforcement, contrasting the limitations of both generalistic human-centred lab studies and application-grounded studies. We explore how XAI can impact the law enforcement environment, drawing from organisational and sociological literature, and from empirical examples from our research at the Netherlands Police Lab for Artificial Intelligence. Our insights show that XAI is no silver bullet towards appropriate reliance but—when designed well—it is one of the factors in the context of a decision that pushes or pulls a person towards appropriate reliance. Ideally, XAI provides considerations to decide to follow, reject, or cross-check an AI's recommendation. We propose a contextual model to understand when and how XAI, along with other contextual factors, can facilitate appropriate reliance on AI recommendations. This model can help to identify individual and task characteristics to assess what factors foster the appropriate reliance on AI recommendations in specific organisational settings.

Funding: The research supervision and writing were funded, in part, by the NWO project ALGOPOL (406.DI.19.011), by the Politie en Wetenschap project Periodieke evaluatie van Mens-AI teams and by a Spinoza grant of the Dutch Research Council (NWO), awarded in 2021 to José van Dijck, Professor of Media and Digital Society at Utrecht University.

M. Bruijnes (✉) · S. Grimmelikhuijsen
Utrecht University School of Governance, Utrecht, Netherlands
e-mail: m.bruijnes@uu.nl; S.G.Grimmelikhuijsen@uu.nl

M. Robeer
Department of Information and Computing Sciences, Utrecht University,
Utrecht, Netherlands
e-mail: M.J.Robeer@uu.nl

1 Introduction

Law enforcement increasingly relies on artificial intelligence (AI) in their day-to-day operations. For instance, police patrols are being managed by predicting high-impact crime¹ and crime reports are automatically labelled to speed up the judgement by police administrators.² The accumulating use of AI as an aid in decision-making in law enforcement has raised important questions, which have received much research attention, on topics such as automated bias³ and a lack of algorithmic transparency.⁴ A less often asked question is whether police officers who work with such algorithms in their daily work can and will trust them, and to what extent they rely on AI recommendations in their decision-making process.⁵ A relatively novel concept contextualises this question: appropriate reliance on explainable artificial intelligence (XAI). XAI has been a major development in computer science. The core idea is that AI applications do not just ‘work’, in the sense that they produce accurate output, but are also explainable to human users. One often-claimed promise of XAI is that it increases the usage of AI recommendations: as users have more insight into why an AI provides a certain recommendation, it will improve the trust that users have, and this will eventually lead to more use.⁶

In this chapter, we problematise that claim by showing that the specific law enforcement context affects the appropriate reliance on XAI recommendations. We argue that there are promising developments, but that convincing empirical evidence for how XAI might work out in complex professional settings is lacking. Experimental studies in the literature tend to use generic (student) samples and settings are non-complex, making these insights transfer poorly to practical settings. These contextual factors are currently ignored in the generic XAI literature, but this needs to be addressed to make the collaboration between users and XAI truly effective.

Second, we challenge the common notion that trustworthiness and usage of AI recommendations are goals in and of themselves. Instead, we argue that XAI ideally leads to *appropriate reliance* on algorithmic advice.⁷ Appropriate reliance is neither stubbornly rejecting, nor complacently accepting AI recommendations. Instead, the XAI provides a set of *considerations* upon which a street-level bureaucrat decides to follow, reject, or cross-check the recommendation. So far, much of the literature on AI and XAI focuses on trust and on ways in which trust can be increased.⁸ However, we argue that trust might not be a goal in and of itself: AI

¹ See Meijer and Wessels (2019).

² See Soares et al. (2023).

³ See Brayne (2017).

⁴ See Mittelstadt et al. (2016).

⁵ See Peeters (2020).

⁶ See Miller (2019).

⁷ See Schemmer et al. (2023).

⁸ See Grimmelikhuijsen and Meijer (2020).

recommendations can be inaccurate, uncertain or unfitting to a particular situation. Instead of increasing ‘blind’ trust in AI recommendations, XAI should contribute to the appropriate reliance on them. In other words, when recommendations are explained—for instance by providing reasons or reliability metrics—it allows law enforcement officers to better assess whether they should rely on a recommendation in a particular situation.⁹

Overall, this chapter highlights—based on conducted empirical research—when law enforcement officers rely on AI recommendations, and how XAI shapes their reliance. The local context in which law enforcement officers operate has important implications for XAI usage. For instance, professional domain knowledge, information complexity and decisional discretion are intertwined with reliance on AI.

We will build this argument as follows. First, we outline current findings in XAI research and how this could be relevant to law enforcement. Second, we discuss literature in organisational and sociological literature to show how technology, and particularly XAI, interacts with its context. Third, we provide real-world empirical examples of XAI-in-context by highlighting research examples from our line of research with graduate students within the Netherlands Police Lab for Artificial Intelligence. These examples are diverse and provide insights into how real police officers deal with (X)AI in their daily practice and the contextual factors that play a role. Finally, based on these research examples, we develop a contextual model to understand when XAI, in concert with other contextual factors, contributes to appropriate reliance on AI recommendations.

2 Current Directions in XAI Research

As highlighted in the Introduction, the use of AI in law enforcement is both promising and contested. On the one hand, it promises a better informed and more efficient use of public resources in law enforcement, on the other hand, current AI systems are opaque and complex for their users.¹⁰ A lack of explainable models and outcomes is at the heart of the debate on the desirability of algorithmic black boxes.¹¹ In this section, we highlight what XAI is, what types of XAI approaches there are, and what empirical studies have found on the effect of XAI in decision-making.

Explanations give decision-makers insight into how advice is constructed and therefore enable informed decision-making. Not all algorithmic explanations, however, provide understanding to a non-technical audience such as law enforcement officers. Doran, Schulz and Besold distinguish between opaque systems, interpretable systems, comprehensible systems, and explainable systems.¹²

⁹ See Alon-Barkat and Busuioc (2023).

¹⁰ See Grimmelikhuijsen and Meijer (2020).

¹¹ See Lepri et al. (2018) and Miller (2019).

¹² See Doran et al. (2017).

Opaque systems do not provide any information about the functioning of the algorithm to the user—but merely provide advice without any explanation or extra information.¹³ *Interpretable* and *comprehensible* systems inform users on which data was used to construct a prediction. Interpretable systems use mathematical terms to do so, while comprehensible systems use human-understandable symbols. However, these systems merely display the information that was used, yet not *why* certain data was used in the recommendation. According to Doran et al., *explainable* systems provide understanding to non-technical audiences and include insight relevant to answering this why question.¹⁴ XAI that supports frontline workers should therefore not (only) display which data was used but also provide the reasons for why the data was used.

The XAI literature broadly distinguishes between two scopes of explanations: *global explanations* that explain the reasoning of the model in general, or *local explanations* that explain algorithmic recommendations in specific situations.¹⁵ Global explanations provide information on the overall model performance, as opposed to showing the reasoning behind a specific recommendation. A global explanation might indicate which variables and data—in general—play a role in the system. Therefore, in the context of law enforcement officers, local explanations may be more appropriate. The tasks of street-level bureaucrats often involve case-by-case judgements and local explanations are useful to provide understanding on such a case-specific basis. Such local explanations can either be embedded proactively in the AI by creating a system that is explainable by design,¹⁶ or extracted retroactively through model-agnostic methods.¹⁷ Still, this does not preclude that global explanations could help them to understand how an algorithmic system works more generally and, for instance, whether the system is appropriate in the context.¹⁸

XAI can take different forms and the type of explanation that is needed depends on the user of the algorithm. In this chapter, we focus on simplified explanations that can be understood by users with domain knowledge, but with no specific technical (AI) knowledge. Such an explanation should, for instance, provide the reasons for what the crucial variable contributed to an algorithmic outcome.¹⁹

Empirical studies investigating the impact of local and global XAI on decision-making aim to ascertain whether providing users with insight into overall AI functioning and its recommendations indeed enhances trust and leads to more effective

¹³ See Burrell (2016).

¹⁴ See Doran et al. (2017).

¹⁵ See Robeer et al. (2023); Adadi and Berrada (2018).

¹⁶ Alvarez-Melis and Jaakkola (2018), p. 8.

¹⁷ See Ribeiro et al. (2016); Lundberg and Lee (2017); van der Waa et al. (2018, 2021).

¹⁸ See Adadi and Berrada (2018); Robeer et al. (2023).

¹⁹ See Friedrich and Zanker (2011); Kizilcec (2016).

utilisation, as posited by Miller.²⁰ Two distinct approaches characterise this inquiry: human-centred or application-grounded.

The first approach uses ‘human-centred’ evaluations to establish general notions of the quality of explainability by using simplified tasks that have little context dependence and typically do not require domain expertise.²¹ In their meta-analysis of nine human-centred evaluation studies on the effect of XAI in human-in-the-loop AI-assisted decision-making, Schemmer et al.²² found that while task performance typically increased with the addition of AI aids, explainability did not always further task performance. The expected utility of explanations, whether confirming sound decisions or raising scepticism towards contradictory evidence, varied across studies. Although XAI often outperformed opaque AI, results were not uniform, with some studies even showing a negative impact. These studies, conducted on participants without domain expertise and oftentimes in an online crowd-sourcing setting (eight out of nine studies were conducted on Amazon Mechanical Turk), had difficulty in understanding *why* XAI did not consistently improve task performance. Proposed explanations included information overload,²³ humans over-relying on the explanations due to task complexity,²⁴ or XAI generally increasing²⁵ or decreasing²⁶ the likelihood of accepting an AI decision—regardless of whether the decision was correct or not.

The second approach in XAI user studies is ‘application-grounded’. These studies evaluate the effect of XAI in a specific application context.²⁷ Such evaluations are typically applied to domain experts, even though they may be non-technical, and shed some light on how different contexts affect the relationship of XAI to decision-making. Haque, Najmul Islam and Mikalef summarised the findings of 58 studies of XAI within 10 application domains, where rather than focusing on the effect of XAI on task performance they took a more fine-grained approach by studying the perceived accuracy, trust, transparency, understandability and fairness of the AI.²⁸ They found that providing decision-makers with information on the context in which the decision is made, historical data, and proper reference of the decision process increased trust in—and perceived transparency of—the AI system, particularly in the healthcare and finance domain. In turn, these beliefs on trust and transparency can affect decision-making behaviour, such as AI adoption and use.

²⁰ See Miller (2019).

²¹ See Doshi-Velez and Kim (2018).

²² See Schemmer et al. (2022).

²³ See Alufaisan et al. (2021).

²⁴ See Liu et al. (2023).

²⁵ See van der Waa et al. (2021).

²⁶ See Fügener et al. (2021).

²⁷ See Doshi-Velez and Kim (2018).

²⁸ See Haque et al. (2023).

Moreover, trust increased with highly confident AI systems,²⁹ while information overload decreased trust.³⁰ Trust was also mediated by whether the participants knew the decision was made by an AI or not, even though this effect was strongest when the decision was positive.³¹ In addition, XAI can also improve the understandability and usability of an AI system.³² Means to do so include training users with the AI before usage³³ and providing clear, concise and comprehensive explanations for non-technical users to prevent information overload.³⁴

While valuable in their exploration of general notions of XAI, human-centred studies often overlook the crucial role of context in the decision-making process. These studies use simplified tasks lacking significant contextual dependencies and frequently involve non-expert participants. Conversely, application-grounded studies, while varying the context wherein the (X)AI operates, tend to focus primarily on evaluating user- and technology-related factors within a context, without explicitly studying the contextual factors and their impact on the effect of XAI on decision-making.

3 A Contextual Approach to XAI

The previous section highlighted the different types of explainable systems and XAI solutions, and how empirical studies currently evaluate the effect of XAI on decision-making. The current literature often studies the technology in isolation or does not explicate the contextual factors affecting the (X)AI-enhanced decision-making process.³⁵ In this section, we therefore discuss XAI along the lines of information sociology and socio-technical studies (STS). Based on this discussion we draw initial lessons on how XAI might (not) contribute to appropriate reliance on AI.

Decades of studies on informatisation and automation in organisations have provided rich insights into how technology and its wider environment interact and shape each other. According to foundational work by Wanda Orlikowski on the duality of technology,³⁶ we should neither see technology as only “hardware” (data, systems, algorithms) nor as purely an outcome of human action. After all, technology has certain properties that influence and limit human actions. At the same time, human actions steer how new technologies further develop. Therefore, she argues, we must see the operation of technology as a product of the interaction between

²⁹ See Ehsan et al. (2021).

³⁰ See Cramer et al. (2008).

³¹ See Bankins et al. (2022).

³² See Branley-Bell et al. (2020).

³³ See Liu et al. (2023).

³⁴ See Hudon et al. (2021).

³⁵ See Miller (2019).

³⁶ See Orlikowski (1992).

humans and technology, which in turn is influenced by the institutional context. In short, (X)AI takes shape and meaning in the institutional, law enforcement, context in which it interacts with police officers. One consequence of this reasoning, however, is that in developing AI we must assume the unpredictability of this technology.³⁷ After all, we do not know how police officers and technology will behave when they interact in practice.

To highlight how the interaction with AI in practice may have unintended effects, we draw from an example from a different research context, namely clinical decision-making.³⁸ An AI-based algorithm was designed to predict the effectiveness of breast cancer drugs. The model was trained on various sources, such as genetic material, clinical signals, and the medical history of patients. The algorithm produced a list of drugs and their expected effectiveness using a numerical score. Based on this list, the physician had to decide which drug was the best choice for their patient. The developers of this system thought this would make the physicians' treatment choices more effective and rational because they would be based on precise and personal data from patients. However, the physicians in practice struggled with the recommendations from the algorithm as they frequently deviated from the professional norms and routines they were familiar with. They had insufficient knowledge of how the algorithm worked, the data that was used and how the output of the algorithm could be interpreted. As a result, the output of the algorithm into an uncertain and ambiguous source of information.

In a similar vein, an ethnographic study by Soares, Grimmelikhuijsen and Meijer highlights that police administrators sometimes simply ignore algorithmic advice.³⁹ In that study, police administrators were supported in their task to classify reports by citizens of suspected online retail fraud. A report could either be a genuine case of online fraud, such as fraudulent web shops that never deliver their products or a 'just' case of mishandling online purchases (a legitimate business that loses a product during delivery). Police administrators were aided by an algorithmic recommendation that classifies a report as "criminal" or "not criminal", based on the name of the retailer and the use of certain words in the report. However, up-close observations of these administrators revealed that they hardly ever used these recommendations and that all reports still underwent manual assessment. Follow-up interviews revealed that one major reason for ignoring algorithmic advice was that the police administrators were unable to understand the rationale and reasoning of algorithmic advice.

This unpredictability and opacity of technology in practice underlines decades of studies of information and communication technologies in the public sector, showing that we need to study organisational practices to understand the unpredictable effects of the use of algorithms.⁴⁰ Meijer and Grimmelikhuijsen use the term

³⁷ See Barley (1986).

³⁸ See Even Chorev (2019).

³⁹ See Soares et al. (2023).

⁴⁰ See Orlikowski (2000).

‘algorithmization’ to describe an organisational process rather than an algorithm as a technological artefact.⁴¹ Building upon older work on informatisation in government by Zuurmond,⁴² Meijer and Grimmelikhuijsen highlight the need to understand new kinds of expertise, the emergence of new information and power relations, and the restructuring of organisational processes.

Here another example helps to show the full extent of this line of reasoning. Lorenz, Meijer and Schupan provide a case study of the introduction of predictive policing at the Berlin police.⁴³ The authors show that a predictive policing algorithm starts to direct the social interactions in the Berlin police: instead of relying on the uncertainty that is fundamental to the professional expertise of police officers, the organisation starts to rely more on the more formalised and seemingly less uncertain recommendations of the algorithm. As a result, decision-making competencies on police patrols are centralised and police professionals on the street become subjected to greater control. Interestingly, these dynamics are not deterministic: in a comparative study with a similar system with the police in Amsterdam, there is much less hierarchical control and more room for professional judgement.⁴⁴

4 Appropriate Reliance on AI in Law Enforcement

The organisational perspective on AI challenges the rationalistic perspective that more data and algorithms will produce more rational decision-making. Instead, we need to think about the processes surrounding the introduction of new technologies to develop ways to strengthen the *appropriate reliance* on AI in law enforcement. We first shortly outline the nascent literature on appropriate reliance in AI research, and then we highlight how the decision-making context complicates overly simple assumptions about ‘correct’ and ‘incorrect’ reliance.

Appropriate reliance on AI has been developed as an alternative to trust in AI.⁴⁵ While trust can be good, especially when an AI recommendation is correct and/or beneficial, AI recommendations are often imperfect and can be wrong. In such cases, we would want people not to trust AI. Explainable AI, then, has been proposed as a solution for humans to better understand the underlying models, variables or reasoning of an algorithm, which could then lead to more trust.⁴⁶ However, based on the concept of appropriate reliance, we expect that a better understanding

⁴¹ See Grimmelikhuijsen and Meijer (2020).

⁴² Zuurmond (1994), pp. 42–48.

⁴³ See Lorenz et al. (2021).

⁴⁴ See Meijer et al. (2021).

⁴⁵ See Benda et al. (2022).

⁴⁶ See Miller (2019).

does not automatically lead to more trust, but instead could lead to a better assessment of and reliance on AI when this is appropriate.⁴⁷

To understand the concept, we briefly explain each term separately. First, what is reliance? Older literature on automation distinguishes reliance from other concepts, such as trust.⁴⁸ Recent research builds on the idea and suggests that reliance, rather than trust, is a better way to conceptualise how humans relate to technology. Based on philosophical work by Baier (1986),⁴⁹ Deley and Dubois highlight that trust “can only exist in relationships where there is a possibility for betrayal”.⁵⁰ In other words, we do not trust technologies in the way we trust people. We rely on technology to succeed, but if technology fails it does not betray us. Furthermore, reliance is often described as a behaviour and thus observable. Scharowski et al. define reliance in the AI advisor context as “*user’s behaviour that follows from the advice of the system*”.⁵¹

Second, when is reliance deemed appropriate? The literature on appropriate reliance distinguishes various ways in which humans can rely on AI recommendations: underreliance, overreliance, and appropriate reliance.⁵² With underreliance, a human user does not follow AI recommendations even when the recommendation is correct. With overreliance, the user follows an incorrect recommendation. This is a useful distinction in some decision contexts when AI advice can be independently and objectively verified as either correct or incorrect. For instance, an AI recommendation which is used to identify a breast cancer tumour on a scan can be correct or incorrect, and a medical doctor can follow this correctly (when the AI is correct) or not (when the AI is incorrect).⁵³

Our main argument in this chapter is that such an approach is not useful in the decision context of law enforcement. Our discussion of this literature shows that in the context of law enforcement appropriate reliance is more complex than currently assumed, due to the nature of their task, as (1) correctness is not an objective feature in many decisions and (2) that correctness is not a binary in many decision contexts.

We will elaborate on these two elements by building on the literature on administrative discretion. The exercise of administrative discretion and direct interactions with clients in decisions is central to the work of street-level bureaucrats, such as law enforcement agents.⁵⁴ Discretion can be described as the freedom that street-level bureaucrats have in applying generic rules to specific cases, in which they have some leeway in determining the quantity and quality of sanctions and rewards when

⁴⁷ See Schemmer et al. (2023).

⁴⁸ See Dzindolet et al. (2003); Lee and See (2004).

⁴⁹ See Baier (1986).

⁵⁰ Deley and Dubois (2020), p. 2.

⁵¹ Scharowski et al. (2022), p. 3.

⁵² See Schemmer et al. (2023).

⁵³ See McKinney et al. (2020).

⁵⁴ Maynard-Moody and Musheno (2000), p. 9.

implementing a policy.⁵⁵ Law enforcement agents exert discretion and often have direct interactions with citizens. This makes the theoretical lens of street-level bureaucracy useful for understanding appropriate reliance in the law enforcement context.

AI affects the administrative discretion of street-level bureaucrats.⁵⁶ Specifically, machine learning approaches not only automate routine administrative tasks but are used to design AI systems that improve the quality of discretionary decision-making of street-level bureaucrats by steering their judgement.⁵⁷ However, the interaction of street-level bureaucrats with AI systems creates new complexities. For instance, a predictive policing system might recommend a police officer survey a certain area, while the police officer—based on experience and local knowledge—estimates that other neighbourhoods have much higher crime risks. Street-level bureaucrats, confronted with such a dilemma, have to decide whether to follow the AI recommendation or their own professional judgement.⁵⁸

As highlighted above, two issues stand out: (1) correctness is not an objective feature in many decisions and (2) correctness is not a binary in many decision contexts. First, predictive systems are probabilistic: they provide chances that some event X might happen. This means that ground truth is often lacking and that only in hindsight the ‘correctness’ of a decision can be assessed. Think of predicting crime or flight routes of suspects. Second, in other instances, in law enforcement work there is often not a binary decision choice, being either correct or incorrect, but there could be multiple decision options with different degrees of correctness, depending on the situation. For instance, an algorithm predicting repeated victimhood of domestic violence can be used to support subsequent actions to protect potential victims: doing nothing because of low risk, providing additional police patrols, assigning protective housing, talking to the perpetrator, etc. Either option can be correct, but the real correctness might only show afterwards, and given the lack of a counterfactual we may never know correctness (if nothing happened, was it due to the taken measures?). Likewise, specific circumstances may cause a police officer to not enforce a rule, which may not be ‘correct’ in the letter of the law but still be correct in terms of individual justice (e.g., waiving a speeding ticket to a woman in labour on a highway).

Therefore, we opt for a definition of appropriate reliance on AI advice that emphasises the context-specific situation of a decision and the possibility that there may be more than one option that can be (in)correct: *Appropriate reliance is to follow up on AI advice while critically evaluating the range of advised decisions or actions by considering their desirability in the specific situation or context.*

⁵⁵ See Tummers and Bekkers (2014).

⁵⁶ See Young et al. (2019).

⁵⁷ See Zouridis et al. (2020).

⁵⁸ See Alon-Barkat and Busuioc (2023) and Selten et al. (2023).

5 XAI in Practice: Empirical Studies at the Police

In this section, we will discuss how XAI could affect appropriate reliance in law enforcement using five empirical research cases. These studies were carried out in the period 2020–2023 by graduate students in the Netherlands Police Lab AI in collaboration with the master program *Public Management* from the *Utrecht University School of Governance*. The studies highlight different aspects of police work but together paint a picture of the contextual factors that affect whether XAI helps to increase appropriate reliance on AI recommendations. Note, however, that these studies did not always explicitly aim to investigate the link between appropriate reliance and XAI, which is indicated in the description of the cases. Two studies take a more qualitative approach, focusing on interviews and experiences of respondents ($n = 41$), and three studies take a quantitative approach involving participants ($n = 437$) in an experiment. These studies report on a *combined sample size of 478 police officers* working in various operational capacities. We describe these case studies next and end each case with the factors contributing to or inhibiting appropriate reliance on (X)AI advice. We summarise all these factors in Table 2 at the end of this section.

5.1 Quantitative Research Cases

Case 1 The thesis by Margot ten Bok asks the question: How does algorithmic transparency affect the perceived moral responsibility of local police officers?⁵⁹ In her study, she investigates a hypothetical digital assistant that searches and ranks information about persons of interest. The AI's task is to report whether citizens or suspects are (rightfully) registered as dangerous, for example, because they are classified as possibly armed. While the AI gives no advice and proposes no actions, it can be argued that it can provide a local explanation of the presented information, as it presents the highest-ranked information based on a clustering algorithm. Note that in this study, appropriate reliance depends on the experimental condition.

To investigate the effect of algorithmic transparency on perceived moral responsibility, a survey experiment was used, with a 2 (AI system makes no mistake/makes a mistake) by 2 (low/high transparency) design. In the low transparency conditions, there is no explanation provided by the AI, while in the high transparency conditions, the system highlights the word used for the rating. In the mistake condition, the explanatory word (e.g., knife) and subsequent classification condition do not match (e.g., the suspect was attacked with a knife, as opposed to the suspect attacked with a knife). The study comprised 153 participants.

⁵⁹ See ten Bok (2022).

The findings indicate that XAI creates a better understanding of the system and increases the chances of detecting nonsensical AI advice, which in turn results in appropriate reliance on that advice. Additionally, it is noted that in practice officers tend to use different information sources to form their judgement and without these sources of information might have to rely on the information of an (X)AI if that is available. Thus, this study suggests that (1) *transparency in AI systems allows for the detection of failures*, such as inappropriate advice, thereby enabling appropriate reliance on the system; and (2) *the absence of other information sources* can create an overreliance on AI advice as there is no other information on which to base a judgement.

Case 2 Gijsbert van Elven asks in his thesis the question: What is the effect of XAI in the form of reliability scores on street-level bureaucrats' trust in algorithmic recommendations, and how does this possible effect differ in the case of congruent or incongruent recommendations?⁶⁰ He studies a hypothetical AI tool that can help police officers estimate the likelihood that a victim who reports to the police will become a victim again in the future. This XAI system utilises a local explanation in the form of a confidence score.

This study utilised a between-subject vignette experiment with a 2 (confidence score present/absent) by 2 (AI advice is congruent/incongruent with the officer's judgement) design. The survey measured the respondents' (n = 160) trust in the AI's advice and whether or not the officer would follow the advice.

The results indicate that congruent advice, which is in line with the professional judgement of the officer, is trusted more and is followed more. Additionally, high confidence scores do not lead to more trust or adoption of the advice, regardless of the congruency of the advice. This suggests that the main factor for relying on AI advice is whether or not the advice is in line with the officer's professional judgement. In other words, they trust and follow AI advice when it confirms what they believe and distrust and do not follow advice that goes against their judgement: a confirmation bias. This would become problematic only in the case of nonsensical AI advice and as such *congruency with own judgement* is considered indicative of over- and underreliance.

Case 3 The thesis by Friso Selten, also published in *Public Administration Review*, asks the question: Does algorithmic advice introduce automation bias and confirmation bias in street-level decision-making, and can explainable AI help mitigate the negative effects of these biases?⁶¹ This thesis investigates a hypothetical AI that predicts the escape route of a criminal directly following a crime. The XAI system provides advice on where to position police officers to best 'fence in' the suspect and offers a local explanation in the form of a rationale for the advice given.

⁶⁰ See van Elven (2023).

⁶¹ See Selten et al. (2023) and Selten (2021).

The study fosters a vignette survey experiment with a 2 (unexplained/explained advice) by 2 (incongruent/congruent advice) design, in addition to a no advice control condition. The congruency vignettes were based on multiple rounds of interviews with police officers and as such reflect common real-world law enforcement knowledge about suspect behaviour. A sample of 124 police officers participated in the study.

This study suggests that police officers are not prone to automation bias, which is an overreliance on AI advice, but rather that they are likely to be subject to confirmation bias, which is an overreliance on their own judgement. Additionally, the findings indicate that XAI might have only a limited effect on police officers in terms of explanations influencing the trust of officers in AI advice. Thus, this study suggests that *congruent AI advice* is a factor that *contributes to overreliance* and that the *proficiency of professionals limits the effects of XAI*.

5.2 Qualitative Research Cases

Case 4 Rosa Massop asks in her thesis the question: What contextual factors influence the occurrence of automation bias and selective follow-up among local police officers?⁶² She investigates this in the context of a hypothetical AI system that offers specialistic knowledge in complex cases that occur so infrequently that officers' general knowledge might not suffice. This AI finds, selects, and prioritises information and presents it in the form of possible actions for officers. In this study, the AI offers no explicit explanation for its advice. However, implicitly the AI does provide a local explanation as it also presents information that is relevant to the case which might be considered as a rationale for the suggested actions.

The method of inquiry is semi-structured interviews about the effects that the contextual factors, which can be seen in Table 1, have on over- and underreliance. Before the interview, participants ($n = 23$) are presented with vignettes that vary whether the AI-suggested actions are in line with the professional judgement and earlier experiences of the officer. The results indicate that overreliance (automation bias) and underreliance (selection bias) can both occur depending on the contextual factors surrounding the decision. The results of this study are summarised in Table 1. A factor that contributes to overreliance (indicated with a plus symbol) is the impact of the decision, whereas the presence of colleagues and experience inhibits overreliance (indicated with a minus). Factors that contribute to underreliance are, for example, uncertainty and complexity surrounding the decision, an experienced decision maker, or workload. Inhibiting underreliance are; the presence of colleagues, and unclear distribution of discretionary space. Note that some results were inconclusive, indicated with a question mark, for example, the effect of workload on overreliance.

⁶² See Massop (2022).

Table 1 Factors contributing to or inhibiting appropriate reliance according to the findings of case 4 by Massop

Factors	Effect on overreliance	Effect on underreliance
<i>Decision</i>		
Uncertainty & complexity	?	+
<i>Decision maker</i>		
Experience	–	+
Workload	?	+
Experienced discretionary space	?	+
<i>Context of the decision</i>		
Presence of colleagues	–	–
Impact of decision	+	?
Distribution of discretionary space between AI and user	?	–

Case 5 The thesis by Mik Collet aims to answer the question: How do police officers experience the use of AI and XAI applications in relation to their discretionary space?⁶³ To this end, he investigates a hypothetical information delivery system that helps officers make complex real-life decisions based on police data and is similar to the AI system in case 4. This AI provides relevant information and gives suggestions for actions. Additionally, it gives local explanations in the form of information relevant to the case and it explains that the conclusion and proposed actions are based on this information.

The design of data gathering was again vignettes followed by semi-structured interviews. The vignettes varied in the explainability of the AI and served as conversation starters and as a framework for the following interview with each of the 18 respondents.

According to this study, AI mostly does not impact the experienced discretionary space of officers, as they critically reflect on multiple information sources. However, XAI mostly positively impacts the experienced discretionary space, as the explanation allows them to critically reflect on the information provided by the AI. More relevant for this chapter, however, are the following findings. Explanations contribute to appropriate reliance as they allow officers to evaluate the reasoning of the AI and judge the appropriateness of the information or advice. Similarly, inhibiting appropriate reliance is a lack of explanation can lead to under-reliance on appropriate information or advice from the AI, as the officer might overly rely on their own experience and judgement. Paradoxically, the presence of explanations can also lead to an overreliance on inappropriate information or advice from the AI, as an officer might be convinced by the XAI’s explanation. This study also suggests a mediator in the form of trust. (Blind) trust in AI, combined with a critical attitude and

⁶³ See Collet (2023).

inquisitive attitude, can lead to overreliance as the explanation can inappropriately convince. However, a lack of trust in an (X)AI can lead to underreliance.

The five case studies together paint a picture of the contextual factors that affect whether XAI helps to increase appropriate reliance on AI recommendations in the police context. These findings are summarised in Table 2, with factors grouped into technology, task and individual.⁶⁴

6 Discussion

In this chapter, we argued that XAI ideally leads to *appropriate reliance* on algorithmic advice. Appropriate reliance is neither stubbornly rejecting, nor complacently accepting AI recommendations. Instead, an XAI provides a set of *considerations* upon which a law enforcement officer decides to follow, reject, or cross-check the AI's recommendation. Based on our research cases we show that current insights from experimental studies with generic student samples and non-complex settings transfer poorly to practical settings. The lack of consensus in the literature about whether (X)AI improves task performance can be better understood by looking into the contextual factors relevant to a particular case. For instance, the presence of an explanation can both contribute to *and* inhibit appropriate reliance. Such a paradox can be understood only by looking into the contextual factors surrounding these decisions: an explanation allows for evaluation of the appropriateness, but an explanation can also be convincing in the wrong direction. Which of these factors takes precedence depends on, yet again, the contextual factors surrounding the decision, for example, the individual's innate trust in AI systems.

We find that XAI is part of a broader information environment, which can steer people towards—or away from—appropriate reliance on AI recommendations. Indeed, XAI may affect appropriate reliance on AI, but reliance is also affected by the task and the individual in a particular context. In this sense, XAI cannot be separated from these other context factors pushing towards and away from appropriate reliance. Additionally, as we have seen in one of the case studies, these contextual factors can act upon each other to strengthen or weaken their effect. Thus, XAI in isolation is insufficient to describe the process leading up to a decision. Figure 1 visualises this idea of how XAI functions in a broader environment.

Figure 1 is best explained by an example, in this case, we use the research example from Selten in which officers used a mock AI tool to predict the flight behaviour of a suspect.⁶⁵ The black-and-white star icon highlights the final decision of a police officer, which is to go to a location to capture the suspect. Figure 1 further shows that a task characteristic, such as high urgency, can reduce reliance, and an individual factor such as strong professional judgements that officers have of suspect

⁶⁴ Awa et al. (2017).

⁶⁵ See Selten (2021).

Table 2 The factors identified in the theses that contribute to or inhibit overreliance, appropriate reliance or underreliance

	Overreliance		Appropriate reliance		Underreliance	
Factors	Contributing	Inhibiting	Contributing	Inhibiting	Contributing	Inhibiting
<i>Technology</i>						
Transparency of AI advice (XAI)	(5)	(1, 5)	(1, 5)			(1, 5)
Lack of information about AI advice (no XAI)	(1)				(5)	
<i>Task</i>						
Experienced discretionary space					(4)	
Workload of professionals during the decision					(4)	
Uncertainty & complexity of the task					(4)	
Impact of the decision	(4)					
Distribution of discretionary space between AI and user						(4)
Presence of colleagues during the decision		(4)				(4)
<i>Individual</i>						
Trust in (X)AI	(5)					
Lack of trust in (X)AI					(5)	
Professional's critical and inquisitive attitude	(5)					
AI advice's congruency with own judgement	(2, 3)			(2)	(2,3)	
Proficiency of professional at making this type of decision		(3, 4)			(4)	(3)

The numbers between brackets refer to the thesis case above

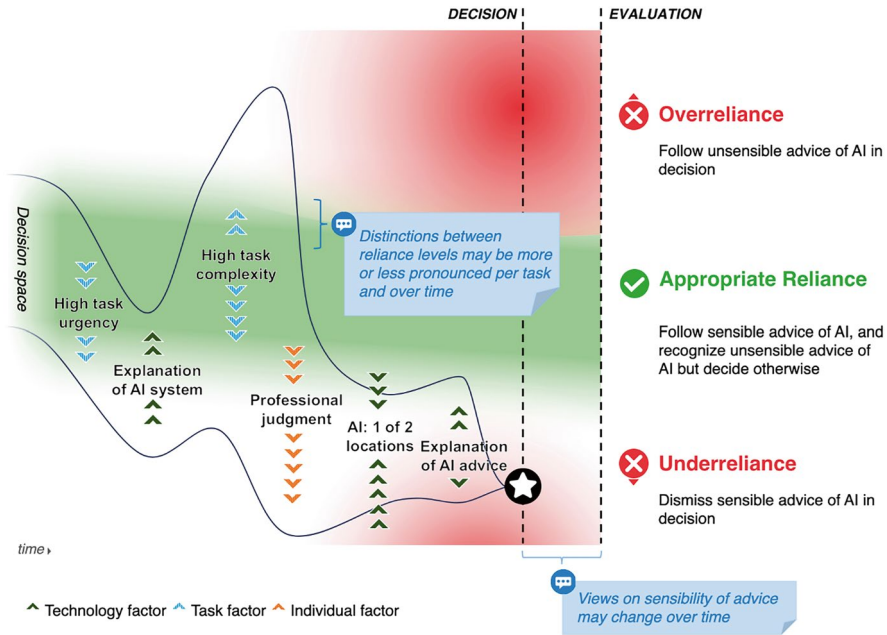


Fig. 1 A heuristic model on how technological, task and individual factors affect appropriate reliance on AI recommendations

flight behaviour, may push towards underreliance. A technological factor, the AI recommendation, only provides two possible flight locations, reducing the discretionary space (visualised by the bandwidth between the two black lines). Within this interplay of contextual factors, a technological factor such as XAI explanation only has a limited effect on the final decision. In Fig. 1, this interplay of the task, individual and technological factors have pushed towards underreliance on the AI's recommendations.

7 Conclusion

Overall, this chapter highlights that XAI is no silver bullet towards appropriate reliance but—when designed well—it is one of the factors in the context of a decision that pushes or pulls a person towards appropriate reliance. At the same time, this implies that we need to allow for sufficient discretion to push and pull towards appropriateness. The local context in which law enforcement officers operate has implications for thinking about the role of XAI in algorithm usage. For instance, professional domain knowledge, information complexity and decisional discretion are intertwined with reliance on AI.

What does this imply for the debate on XAI? First, we find that the influence of XAI—while important—should not be overstated and that the debate on the appropriate reliance on AI recommendations should be broadened to non-technological factors, such as those related to the task and individual environment. In research on AI recommendations, we should not take the technology as the point of departure, but the task it is designed to improve, and the individual who is intended to use the technology.

In this sense, our conclusion shows the value of placing the role of XAI in a research tradition that studies organisational practice to understand the interaction between users and technology.⁶⁶ The use of XAI in practice is often opaque and unpredictable at the outset of its deployment. When we assume such unpredictability, we can gain a much deeper understanding of how XAI unfolds by studying organisational practices.

This implies that future research on Explainable AI needs to take the organisational context into account. For instance, a country comparison between a police department in Germany and the Netherlands showed that management styles influence how police officers view and use predictive policing systems.⁶⁷ XAI could be seen as one element in this context that affects if and how police officers rely on such predictive systems.

Future research could further extend this kind of knowledge by extending to other domains than law enforcement, for instance, medical decisions⁶⁸ and public service delivery.⁶⁹ The heuristic model (Fig. 1) is flexible and could help future studies to identify individual and task characteristics to assess what factors foster the appropriate reliance on AI recommendations in organisational settings.

Acknowledgements We wish to express our gratitude and acknowledge the efforts of the students who carried out their thesis project in the Master Public Management at the department of Public Administration and Organizational Science (Utrecht University): Margot ten Bok, Rosa Massop, Friso Selden, Mik Collet and Gijsbert van Elven. We are also grateful to the Netherlands Police and the Nationale Politielab AI for facilitating and supporting the research efforts of these students.

References

- Adadi A, Berrada M (2018) Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). *IEEE Access* 6:52138–52160. <https://doi.org/10.1109/ACCESS.2018.2870052>
- Alon-Barkat S, Busuioc M (2023) Human–AI interactions in public sector decision making: “automation bias” and “selective adherence” to algorithmic advice. *J Public Adm Res Theory* 33(1):153–169. <https://doi.org/10.1093/jopart/muac007>

⁶⁶ See Orlikowski (2000).

⁶⁷ See Meijer et al. (2021).

⁶⁸ See Even Chorev (2019).

⁶⁹ Alon-Barkat and Busuioc (2023).

- Alufaisan Y, Marusich LR, Bakdash JZ, Zhou Y, Kantarcioglu M (2021) Does explainable artificial intelligence improve human decision-making? *Proc AAAI Conf Artif Intell* 35(8):(Article 8) 6618–6626. <https://doi.org/10.1609/aaai.v35i8.16819>
- Alvarez-Melis D, Jaakkola TS (2018) Towards robust interpretability with self-explaining neural networks. *Adv Neural Inf Proces Syst* 31. <https://proceedings.neurips.cc/paper/2018/hash/3e9f0fc9b2f89e043bc6233994dfcf76-Abstract.html>
- Awa HO, Ojiabo OU, Orokor LE (2017) Integrated technology-organization-environment (T-O-E) taxonomies for technology adoption. *J Enterp Inf Manag* 30(6):893–921. <https://doi.org/10.1108/JEIM-03-2016-0079>
- Baier A (1986) Trust and antitrust. *Ethics* 96(2):231–260. <https://doi.org/10.1086/292745>
- Bankins S, Formosa P, Griep Y, Richards D (2022) AI decision making with dignity? Contrasting workers' justice perceptions of human and AI decision making in a human resource management context. *Inf Syst Front* 24(3):857–875. <https://doi.org/10.1007/s10796-021-10223-8>
- Barley SR (1986) Technology as an occasion for structuring: evidence from observations of CT scanners and the social order of radiology departments. *Adm Sci Q* 31(1):78–108. <https://doi.org/10.2307/2392767>
- Benda NC, Novak LL, Reale C, Ancker JS (2022) Trust in AI: why we should be designing for APPROPRIATE reliance. *J Am Med Inform Assoc* 29(1):207–212. <https://doi.org/10.1093/jamia/ocab238>
- Branley-Bell D, Whitworth R, Coventry L (2020) User trust and understanding of explainable AI: exploring algorithm visualisations and user biases. In: Kurosu M (ed) *Human-computer interaction. Human values and quality of life*. Springer International Publishing, pp 382–399. https://doi.org/10.1007/978-3-030-49065-2_27
- Brayne S (2017) Big data surveillance: the case of policing. *Am Sociol Rev* 82(5):977–1008. <https://doi.org/10.1177/0003122417725865>
- Burrell J (2016) How the machine 'thinks': understanding opacity in machine learning algorithms. *Big Data Soc* 3(1):2053951715622512. <https://doi.org/10.1177/2053951715622512>
- Collet M (2023) (X)AI en de politieagent: intelligente assistent of bedreigende concurrent? Master Thesis, Utrecht University
- Cramer H, Evers V, Ramlal S, van Someren M, Rutledge L, Stash N, Aroyo L, Wielinga B (2008) The effects of transparency on trust in and acceptance of a content-based art recommender. *User Model User-Adap Inter* 18(5):455–496. <https://doi.org/10.1007/s11257-008-9051-3>
- Deley T, Dubois E (2020). Assessing trust versus reliance for technology platforms by systematic literature review. *Soc Media Soc* 6(2):2056305120913883. <https://doi.org/10.1177/2056305120913883>
- Doran D, Schulz S, Besold TR (2017) What does explainable AI really mean? A new conceptualization of perspectives (arXiv:1710.00794). *arXiv*. <https://doi.org/10.48550/arXiv.1710.00794>
- Doshi-Velez F, Kim B (2018) Considerations for evaluation and generalization in interpretable machine learning. In: Escalante HJ, Escalera S, Guyon I, Baró X, Güçlütürk Y, Güçlü U, van Gerven M (eds) *Explainable and interpretable models in computer vision and machine learning*. Springer International Publishing, pp 3–17. https://doi.org/10.1007/978-3-319-98131-4_1
- Dzindolet MT, Peterson SA, Pomranky RA, Pierce LG, Beck HP (2003) The role of trust in automation reliance. *Int J Human-Comput Stud* 58(6):697–718. [https://doi.org/10.1016/S1071-5819\(03\)00038-7](https://doi.org/10.1016/S1071-5819(03)00038-7)
- Ehsan U, Wintersberger P, Liao QV, Mara M, Streit M, Wachter S, Riener A, Riedl MO (2021) Operationalizing human-Centered perspectives in explainable AI. In: Kitamura Y, Quigley A (eds) *Extended abstracts of the 2021 CHI conference on human factors in computing systems*, Article 94. Association for Computing Machinery, pp 1–6. <https://doi.org/10.1145/3411763.3441342>
- Even Chorev N (2019) Data ambiguity and clinical decision making: a qualitative case study of the use of predictive information technologies in a personalized cancer clinical trial. *Health Informatics J* 25(3):500–510. <https://doi.org/10.1177/1460458219827355>

- Friedrich G, Zanker M (2011) A taxonomy for generating explanations in recommender systems. *AI Mag* 32(3):90–98. <https://doi.org/10.1609/aimag.v32i3.2365>
- Fügener A, Grahl J, Gupta A, Ketter W (2021) Will humans-in-the-loop become Borgs? Merits and pitfalls of working with AI. *Manage Inform Syst Q* 45(3):1527–1556. <https://papers.ssrn.com/abstract=3879937>
- Grimmelikhuijsen S, Meijer A (2020) Responsible and accountable algorithmization: how to generate citizen trust in governmental usage of algorithms. In: Schuilenburg M, Peeters R (eds) *The algorithmic society: technology, power, and knowledge*. Routledge, pp 53–66
- Haque AB, Islam AKMN, Mikalef P (2023) Explainable artificial intelligence (XAI) from a user perspective: a synthesis of prior literature and problematizing avenues for future research. *Technol Forecast Soc Chang* 186:122120. <https://doi.org/10.1016/j.techfore.2022.122120>
- Hudon A, Demazure T, Karran A, Léger P-M, Sénécal S (2021) Explainable artificial intelligence (XAI): how the visualization of AI predictions affects user cognitive load and confidence. In: Davis FD, Riedl R, vom Brocke J, Léger P-M, Randolph AB, Müller-Putz G (eds) *Information systems and neuroscience*. Springer International Publishing, pp 237–246. https://doi.org/10.1007/978-3-030-88900-5_27
- Kizilcec RF (2016) How much information? Effects of transparency on trust in an algorithmic interface. In: *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 2390–2395. <https://doi.org/10.1145/2858036.2858402>
- Lee JD, See KA (2004) Trust in automation: designing for appropriate reliance. *Hum Factors* 46(1):50–80. https://doi.org/10.1518/hfes.46.1.50_3039
- Lepri B, Oliver N, Letouzé E, Pentland A, Vinck P (2018) Fair, transparent, and accountable algorithmic decision-making processes. *Philosophy Technol* 31(4):611–627. <https://doi.org/10.1007/s13347-017-0279-x>
- Liu G, Zhang J, Chan AB, Hsiao J (2023) Human attention-guided explainable AI for object detection. In: Goldwater M, Anggoro FK, Hayes BK, Ong DC (eds) *Proceedings of the 45th Annual Conference of the Cognitive Science Society*, pp 2573–2580. <https://escholarship.org/uc/item/9r53b44n>
- Lorenz L, Meijer A, Schuppan T (2021) The algocracy as a new ideal type for government organizations: predictive policing in Berlin as an empirical case. *Inform Polity* 26(1):71–86. <https://doi.org/10.3233/IP-200279>
- Lundberg SM, Lee S-I (2017). A unified approach to interpreting model predictions. In: Guyon I, Von Luxburg U, Bengio S, Wallach H, Fergus R, Vishwanathan S, Garnett R (eds) *Advances in neural information processing systems* 30 (NIPS 2017). Neural Information Processing Systems Foundation, pp 4765–4774. <https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html>
- Massop R (2022) Wie bepaalt: het systeem of de politieagent? Een kwalitatief vignettenonderzoek naar automatiseringsbias en selectieve opvolging bij het werken met AI toepassingen bij de politie. Master Thesis, Utrecht University. <https://studenttheses.uu.nl/handle/20.500.12932/43262>
- Maynard-Moody S, Musheno M (2000) State agent or citizen agent: two narratives of discretion. *J Public Adm Res Theory* 10(2):329–358. <https://doi.org/10.1093/oxfordjournals.jpart.a024272>
- McKinney SM, Sieniek M, Godbole V, Godwin J, Antropova N, Ashrafian H, Back T, Chesus M, Corrado GS, Darzi A, Etemadi M, Garcia-Vicente F, Gilbert FJ, Halling-Brown M, Hassabis D, Jansen S, Karthikesalingam A, Kelly CJ, King D et al (2020) International evaluation of an AI system for breast cancer screening. *Nature* 577(7788):89–94. <https://doi.org/10.1038/s41586-019-1799-6>
- Meijer A, Wessels M (2019) Predictive policing: review of benefits and drawbacks. *Int J Public Adm* 42(12):1031–1039. <https://doi.org/10.1080/01900692.2019.1575664>
- Meijer A, Lorenz L, Wessels M (2021) Algorithmization of bureaucratic organizations: using a practice lens to study how context shapes predictive policing systems. *Public Adm Rev* 81(5):837–846. <https://doi.org/10.1111/puar.13391>
- Miller T (2019) Explanation in artificial intelligence: insights from the social sciences. *Artif Intell* 267:1–38. <https://doi.org/10.1016/j.artint.2018.07.007>

- Mittelstadt BD, Allo P, Taddeo M, Wachter S, Floridi L (2016) The ethics of algorithms: mapping the debate. *Big Data Soc* 3(2):2053951716679679. <https://doi.org/10.1177/2053951716679679>
- Orlikowski WJ (1992) The duality of technology: rethinking the concept of technology in organizations. *Organ Sci* 3(3):398–427. <https://doi.org/10.1287/orsc.3.3.398>
- Orlikowski WJ (2000) Using technology and constituting structures: a practice lens for studying technology in organizations. *Organ Sci* 11(4):404–428. <https://doi.org/10.1287/orsc.11.4.404.14600>
- Peeters R (2020) The agency of algorithms: understanding human-algorithm interaction in administrative decision-making. *Inform Polity* 25(4):507–522. <https://doi.org/10.3233/IP-200253>
- Ribeiro MT, Singh S, Guestrin C (2016) “Why should I trust you?”: explaining the predictions of any classifier. In: *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp 1135–1144. <https://doi.org/10.1145/2939672.2939778>
- Robeer M, Bex F, Feelders A, Prakken H (2023) Explaining model behavior with global causal analysis. In: Longo L (ed) *Explainable artificial intelligence*. Springer Nature Switzerland, pp 299–323. https://doi.org/10.1007/978-3-031-44064-9_17
- Scharowski N, Perrig SAC, von Felten N, Brühlmann F (2022) Trust and reliance in XAI -- distinguishing between attitudinal and behavioral measures (arXiv:2203.12318). arXiv. <https://doi.org/10.48550/arXiv.2203.12318>
- Schemmer M, Hemmer P, Kühl N, Benz C, Satzger G (2022) Should I follow AI-based advice? Measuring appropriate reliance in human-AI decision-making (arXiv:2204.06916). arXiv. <https://doi.org/10.48550/arXiv.2204.06916>
- Schemmer M, Kühl N, Benz C, Bartos A, Satzger G (2023) Appropriate reliance on AI advice: conceptualization and the effect of explanations. In: *Proceedings of the 28th International Conference on Intelligent User Interfaces*. Association for Computing Machinery, pp. 410–422. <https://doi.org/10.1145/3581641.3584066>
- Selten FJ (2021) Can explainable AI mitigate decision-making errors induced by algorithms in street-level police work? An experiment. Master Thesis, Utrecht University. <https://studenttheses.uu.nl/handle/20.500.12932/40834>
- Selten FJ, Robeer M, Grimmelikhuijsen S (2023) ‘Just like I thought’: street-level bureaucrats trust AI recommendations if they confirm their professional judgment. *Public Adm Rev* 83(2):263–278. <https://doi.org/10.1111/puar.13602>
- Soares C, Grimmelikhuijsen S, Meijer A (2023) Screen-level bureaucrats in the age of algorithms. An ethnographic study of algorithmically supported public service workers in The Netherlands police. *Inform Polity* 29(3):277–292. <https://doi.org/10.3233/IP-220070>
- ten Bok M (2022) Agent & Algoritme. Master Thesis, Utrecht University. <https://studenttheses.uu.nl/handle/20.500.12932/42031>
- Tummers L, Bekkers V (2014) Policy implementation, street-level bureaucracy, and the importance of discretion. *Public Manag Rev* 16(4):527–547. <https://doi.org/10.1080/14719037.2013.841978>
- van der Waa J, Robeer M, van Diggelen J, Brinkhuis M, Neerinx M (2018) Contrastive Explanations with Local Foil Trees (arXiv:1806.07470). arXiv. <https://doi.org/10.48550/arXiv.1806.07470>
- van der Waa J, Nieuwburg E, Cremers A, Neerinx M (2021) Evaluating XAI: a comparison of rule-based and example-based explanations. *Artif Intell* 291:103404. <https://doi.org/10.1016/j.artint.2020.103404>
- van Elven G (2023) Indicator versus intuïtie: Een experimenteel onderzoek naar het effect van betrouwbaarheidsindicatoren op het vertrouwen van politieagenten in algoritmische aanbevelingen. Master Thesis, Utrecht University
- Young MM, Bullock JB, Leczy JD (2019) Artificial discretion as a tool of governance: a framework for understanding the impact of artificial intelligence on public administration. *Perspect Public Manage Gov* 2(4):301–313. <https://doi.org/10.1093/ppmgov/gvz014>

- Zouridis S, van Eck M, Bovens MAP (2020) Automated discretion. In: Evans T, Hupe P (eds) Discretion and the quest for controlled freedom. Springer International Publishing, pp 313–329. https://doi.org/10.1007/978-3-030-19566-3_20
- Zuurmond A (1994) De infocratie: Een theoretische en empirische heroriëntatie op Weber's ideaal-type in het informatietijdperk = The infocracy. Phaedrus

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.



European Digital Infrastructure Consortium (EDIC): A New Governance Framework for the European Blockchain Services Infrastructure (EBSI)



Evrin Tan and Daniël Du Seuil

Abstract This chapter explores the inception of the European Digital Infrastructure Consortium (EDIC), a new institutional model for the governance of the European Blockchain Services Infrastructure (EBSI), a pan-European digital framework for cross-border public services. Evolving from the previous European Blockchain Partnership (EBP), the shift to Europeum was driven by technological progress, organizational requirements, and regulatory framework adjustments. The chapter first examines the EDIC as an innovative institutional framework proposed by the Digital Decade Policy Programme 2030 to oversee multi-country digital projects. Subsequently, it delves into the formation process of Europeum's governance by detailing ongoing organizational decisions. Additionally, the chapter navigates through the governance challenges linked to this cross-national initiative via desk research of working documents and expert interviews, exploring the motivations and rationale behind establishing a pioneering cross-national state-led governance model for a blockchain-based system.

1 Introduction

Blockchain technology and related concepts such as self-sovereign identity systems, verifiable credentials, and decentralized identifiers are progressively reshaping information management systems across various sectors within the public domain.¹ A noteworthy global initiative harnessing these innovations is the European Blockchain Services Infrastructure (EBSI), a collaborative initiative between the

¹ See Tan et al. (2022); Benchaya Gans et al. (2022); Cagigas et al. (2021).

E. Tan (✉)
Public Governance Institute, KU Leuven, Leuven, Belgium
e-mail: evrim.tan@kuleuven.be

D. Du Seuil
HoWest Hogeschool West-Vlaanderen, Kortrijk, Belgium
e-mail: daniel.du.seuil@howest.be

European Commission (EC) and the European Blockchain Partnership (EBP). The EBP established in 2018, brings together 27 EU member states, along with Norway and Liechtenstein, to develop an EU strategy on blockchain and build a blockchain infrastructure, i.e. EBSI, for public services.² EBSI's primary objective is to establish a pan-European infrastructure designed to facilitate cross-border services for public institutions, businesses, citizens, and their interconnected ecosystems. This infrastructure aims to enhance the verification of information and foster trustworthiness in services.

Since 2020, EBSI has undertaken the deployment of a distributed network of nodes spanning Europe. It has been actively engaged in the development and testing of various use cases across domains such as education and social security.³ These use cases aim to leverage the EBSI blockchain and compatible decentralized technologies to enhance the delivery of public services in cross-border settings.

Around the same time, the EC launched its policy strategy 'Digital Decade', a strategy document that aims to transform the digital public services of the EU countries until 2030. In an era characterized by the rapid expansion of digital technologies and their transformative impact on societies and economies, the EU acknowledges the need for innovative institutional arrangements to govern digital infrastructures effectively. One such response is the establishment of European Digital Infrastructure Consortium (EDIC), a novel institutional instrument within the EU's governance framework designed to coordinate the implementation of multi-country digital projects. EDIC represents a strategic development in the cross-national governance of digital infrastructures, seeking to harness the collective capabilities of Member States (MS) and other entities for the advancement of digitalization across Europe.

In 2023, EBP marked a significant milestone as being one of the first networks applied to establish an EDIC for governing EBSI. This EDIC was formally established in May 2024 under the name of *Europeum*. This chapter introduces this novel institutional instrument of the EU for digital infrastructures and investigates the motivation and influences shaping the governance structure of EDIC. Drawing on desk research with working documents and a series of expert interviews with the task force assigned to establish the statute of the *Europeum*, we explore the task force members' perceptions and expectations regarding the main technical, political, financial, and legal challenges for the establishment and the continuity of the EBSI blockchain and its governance under EBP and *Europeum*. Furthermore, we investigate the task force members' views on the future trajectory of this digital infrastructure and how they anticipate its governance structure evolving to influence global blockchain governance across public and private sectors. Specifically, the chapter addresses key questions: (1) What is the proposed governance structure of *Europeum*? (2) Which governance challenges influenced the shift from EBP to

²See the announcement at EC (June 2022): <https://digital-strategy.ec.europa.eu/en/policies/blockchain-partnership>.

³See Tan et al. (2023).

Europeum? (3) What are the perceived challenges in Europeum's governance formation? The discussion section further examines these findings through the lenses of isomorphism, organizational transformation, and collaborative governance theories, offering deeper insights into the governance challenges faced by cross-national blockchain-based infrastructures.

2 European Digital Infrastructure Consortium (EDIC) as a New Cross-National Governance Model

An EDIC represents the institutional response of the Digital Decade Programme (DEP),⁴ which aims to accelerate the digital transformation of the Union and its member countries. Entering into force in November 2022, the strategy document⁵ sets the legal and institutional conditions for establishing and functioning of an EDIC to address the challenges in implementing multi-country digital infrastructures and projects. EDIC allows multi-country projects to gain a legal capacity to operate in and across MS, allowing them to acquire, own, and dispose movable, immovable and intellectual property within the area of operation.⁶ An EDIC, hosted in a participating Member State, must be recognized by all MS as either an international body or an international organization.⁷ To establish an EDIC, interested MS submit an application to the EC, including the proposed statutes of the EDIC, a technical description of the multi-country project, and a declaration by the host Member State recognizing the EDIC as an international body or organization. The EC evaluates applications based on set criteria and may establish or reject an EDIC through interservice consultation with different levels. Application requires at least three contributing MS with voting rights granted to those providing financial or non-financial contributions. Other MS may join later, while those without contributions can participate as observers. Membership may extend to entities beyond MS, with MS retaining majority voting rights regardless of external contributions.⁸ An EDIC may deploy joint infrastructure, deliver related services, and act as a networking platform to bring together public and private stakeholders. As of June 2024, there are three EDICs established by the MS on language technologies, digital twins, and blockchain.⁹

As a legal entity, an EDIC is established through the Commission decision at the request of a group of MS. This decision grants the EDIC full legal capacity, recognising its existence and authority across all MS. Importantly, the EDIC assumes

⁴EU Decision 2022/2481.

⁵See the announcement at EC (June 2024): <https://digital-strategy.ec.europa.eu/en/policies/edic>.

⁶EU Decision 2022/2481, Art.14.

⁷EU Decision 2022/2481, Art 1(45).

⁸EU Decision 2022/2481, Art.14 and Art.15.

⁹EC (June 2024): <https://digital-strategy.ec.europa.eu/en/policies/edic>.

liability for its debts, signifying its legal independence and financial responsibility. In this regard, the EDIC shares legal characteristics with other European multi-country organizational forms, such as the European Cooperative Society (ECS) or European Economic Interest Grouping (EEIG), designed to facilitate collaboration and cooperation among companies and cooperatives in cross-border activities. However, unlike these organizational forms, EDICs aim to bring together MS rather than the corporations to deploy joint infrastructure for service delivery under cross-border conditions.

As an organizational form, the EDIC most closely resembles the European Research Infrastructure Consortium (ERIC), a European-level legal entity for establishing and operating new or existing research infrastructures on a non-economic basis. However, unlike ERICs, EDICs are governed by Member State representatives, and their budget may include service fees in addition to Member State or EU contributions. In this sense, EDICs appear closer to an International Organization (IO) but also possess characteristics of a multinational corporation.

The primary sources of financing for EDIC are rooted in the commitments made by its member states and grants obtained from directly managed EU instruments. While the former underscores the cooperative financial engagement of MS, the latter opens avenues for EDICs to seek funding from a diverse array of centrally managed EU programs. These programs include the Digital Europe Programme (DEP), the Connecting Europe Facility (CEF), Horizon Europe, the Recovery and Resilience Facility (RRF), and funds under shared management.

An EDIC's governance framework is distinctly characterized by the central role of MS that provide either financial or in-kind contributions. The statute governing the EDIC provides the framework for its internal organization, which can adapt to the specific requirements and preferences of its members. Financially contributing MS wield voting rights within the EDIC's decision-making processes, underscoring their influence over the consortium's activities and strategic direction. MS that do not provide financial contributions still have the opportunity to engage with EDIC, albeit as observers without voting rights. This duality in membership status reflects the diverse participation dynamics within the EDIC, accommodating the interests and capacities of various MS.

While an EDIC is designed to operate as a legal entity with full legal capacity, it does not assume the status of a Union body. Consequently, the EC does not hold voting rights within the EDIC's governance structure, reflecting the distinct character of this institutional instrument. However, the EC representatives still retain a significant role in the EDIC's governance, particularly in matters pertaining to the strategic orientation of the operations. The EC's engagement extends to the deliberations of the assembly through an advisory board within the governance structure. Furthermore, the EC representatives may exercise a veto right for decisions related to actions financed by the EU programmes. This oversight underscores the EC's pivotal role in safeguarding the alignment of EDICs' activities with broader EU policies and objectives.

3 From the European Blockchain Partnership to Europeum

The governance formation of European Blockchain Services Infrastructure (EBSI) began with the European Blockchain Partnership (EBP), an informal agreement that brought together MS coordinated by the EC. Since its inception in 2020, EBSI has embarked on the deployment of a distributed network of nodes across Europe, concurrently developing and testing use cases through the “Early Adopters Programme” (EAP) across domains such as education and social security. These use cases draw extensively from self-sovereign identity systems, verifiable credentials, and decentralized identifiers, all of which constitute critical components of the EBSI ecosystem.¹⁰ Several pilot actions, projects, and initiatives have been launched, creating a network of over 400 participating organizations.¹¹ Within this network, there’s an increasing number of nodes governed by authorized organizations that contribute to the EBSI network. These nodes, numbering 39 as of 2023, where 26 of them acting as validator nodes, serve as the technical backbone of EBSI.¹² Node operators are selected based on their operational capacity and their adherence to the specific security and availability standards of the EBSI governance rules. They create electronic ledgers that are jointly validated and updated in a synchronized manner, ensuring data integrity and reliability.¹³

The initial governance of the EBSI ecosystem was established through a memorandum of understanding between EBP partners and the EC with an indication of tasks, obligations and funding. On the EU side, the governance of the EBSI ecosystem contained a collaboration between the European Commission Director Generals of Communication, Networks, and Technology (DG CNECT) and of Informatics (DIGIT). Until the formation of EDIC, DG CNECT acts as the coordinating body as the solution owner for financial and policy orientations, and DIGIT focuses on the service provision concerning technical implementation. On the EBP side, technical, policy and use case groups formed by the country representatives of the MS meet periodically to deliberate on the development of the technical and policy capacities to coordinate the actions across MS.

The development of EBSI has gone through incremental stages developing an ecosystem of actors and use cases. Three core initiatives have been influential in this development stage. First, through the financing of the Connected Europe Facility and Digital Europe Programme, with a budget of approximately 10 million Euro from the EU budget, several multi-partner projects were funded to develop the technical and institutional capabilities of European actors in developing use cases, node

¹⁰ See Tan et al. (2023).

¹¹ See EBSI (2024b): <https://ec.europa.eu/digital-building-blocks/sites/display/EBSI/Home>.

¹² See EBSI (2024c): <https://ec.europa.eu/digital-building-blocks/sites/display/EBSI/Node+Operator+Map>.

¹³ See Tan et al. (2023).

operations, and technical standards (e.g. EBSI-Vector, EBSI-NE, EBSI-Standard, etc).¹⁴

Secondly, through the initiation of the Early Adopter's Programme in 2021, local projects and proof-of-concepts were supported to upscale to cross-border pilots compatible with EBSI blockchain. Especially, in the accreditation of academic documents and verification of identities, Early Adopter's Programme has supported 21 projects from 18 European countries.¹⁵

Thirdly, EBP has initiated different projects to create an operational ecosystem for EBSI. These initiatives have focused on the development of technical standards concerning node operators, smart contracts, digital wallets (e.g. EBSI conformance testing), application service providers (ASPs), infra-services and tools, improvement proposals, and also collaborating with other initiatives and institutions such as EU Digital Identity Wallet (EUDI), European Social Security Pass (ESSPASS), European Intellectual Property Office (EUIPO), and Europass.

Once fully operational, the EBSI can be used for the secure and trusted exchange of certificates, information, or data across borders and within the participating countries. Moreover, the EBSI services can be used with a digital wallet operated through a smartphone, allowing citizens to control the usage of their personal data by third parties while preserving their privacy. This will be done in conformity with the proposed regulation of EIDAS 2.0, establishing a framework for a European Digital Identity (EUDI) and the EUDI wallet.

Throughout this operational evolution, EBP has provided a platform for MS and the EC to develop the policies and technical standards for the formation of EBSI as a European infrastructure of cross-border services. However, neither the CNECT nor the EBP has the mandate to financially support and operate such a technical infrastructure in a full production environment for cross-border public services. Therefore, establishment of an EDIC has been proposed as the next institutional phase for the governance of EBSI.

In June 2023, Belgium, Italy, Portugal, Croatia, and Slovenia, joined later by Luxembourg, Greece, Cyprus, Poland, and Romania, agreed to establish Europeum as an EDIC to replace EBP, with Belgium becoming the hosting country. Europeum is set to operate under Belgian law, having the legal seat in Brussels, and supporting the development, deployment, and operation of EBSI. According to the draft statute, the Europeum members will be responsible for maintaining blockchain nodes, supporting EBSI application service providers, and establishing partnerships for new use cases. It will seek regulatory compliance across MS and engage with relevant networks and organizations in technology interoperability, research, innovation support, skills development, and collaboration with technological initiatives.

The draft statute of the Europeum sets the main responsibilities of Europeum as follows:

¹⁴ See EBSI (2024a) for further information: <https://hub.ebsi.eu/>.

¹⁵ See EBSI (2024c) for further information: <https://ec.europa.eu/digital-building-blocks/sites/display/EBSI/Verifiable+Credentials+Success+Stories>.

- *Establishing EBSI for cross-border services:* The primary task of Europeum is to establish EBSI and operate it to provide EU-wide cross-border services, especially in the domain of public services. The central goal is to enhance trust and cyber resilience in the delivery of these services while ensuring compatibility with the European Digital Identity trust model.
- *Fostering European digital identity integration:* Europeum also explores the potential of EBSI in contributing to use cases associated with the implementation of the European Digital Identity framework. This exploration includes harnessing trusted electronic ledgers and verifiable credentials to bolster the digital identity ecosystem.
- *Supporting cross-border cooperation and innovation:* Europeum promotes cross-border cooperation among public authorities in the realm of decentralized technologies. It actively facilitates the interoperability of EBSI solutions with other cutting-edge technologies, operating at the levels of protocols, smart contracts, and applications. This commitment is instrumental in creating a conducive environment for innovation.

To finalize the institutional formation of Europeum in early 2024, and to facilitate the transition process, a task force was established by the representatives of founding members of EDIC. This task force is to lead the governance structure formation and establish formal rules and statuses. Subsequent sections will provide a detailed examination of Europeum's governance formation based on the working documents shared by the task force and address perceived governance challenges as outlined by the task force members.

4 Governance of Europeum

The EU decision¹⁶ regarding the establishment of the Digital Decade Policy Programme 2030, obliges every EDIC to comprise two main bodies: an assembly of members and a director. The assembly of members consists of MS, the EC, and international organizations of European interest, and public or private entities.¹⁷ The assembly serves as the primary decision-making body with authority over budget adoption and other significant matters. The director, appointed by the assembly of members, acts as the executive body and legal representative of the EDIC. While the EC participates in the assembly's deliberations, it does not possess voting rights, except in cases where centrally managed Union programmes contribute to the project, granting the EC a right of veto on decisions pertaining solely to actions financed under such programs.

¹⁶ EU Decision 2022/248.

¹⁷ *Ibid*, Art. 15(4).

While the EU Decision 2022/248 sets the general premises of the EDIC governance, the European Digital Infrastructure Consortium User Guide¹⁸ includes further descriptive clauses in the governance formation of the EDICs.¹⁹ Accordingly, while the assembly of members and the Director are mentioned as ‘mandatory elements’, EDICs have the flexibility to incorporate non-mandatory governance elements such as advisory boards tailored to its specific needs. The guide mentions two types of such advisory boards: a Strategic Orientation Committee and an Advisory Board.

The Strategic Orientation Committee is described as a supportive body that would gather representatives of different categories of members and would provide a more flexible platform than the Assembly of Members for discussing strategic matters crucial to the EDIC’s mission. Although its recommendations to the Director are expected to be non-binding, the Assembly of Members may choose to formalize its guidance.

The Advisory Board, on the other hand, is envisaged as a body to offer insights from external organizations and experts, particularly from user communities not eligible for membership. Depending on the needs, other types of bodies may be established for EDICs, but to avoid complications in governance, these bodies are suggested not to have decision-making competences. While not mandated, following the creation of an EDIC, it is recommended that the Assembly to adopt implementing rules outlining internal organizational procedures comprehensively. These rules, covering financial, procurement, voting, and hiring processes, can be periodically updated to ensure operational efficiency without necessitating statutory amendments.

In what follows, we synthesize working documents developed by the task force regarding the governance structure of EDIC. These documents are not finalized, so it’s possible that the governance structure could undergo further changes once the institutionalization of EDIC is completed. However, the following governance dimensions are currently approved by the task force and Fig. 1 illustrates the governance model envisioned for Europeum.

The proposed governance structure of EDIC envisions an interactive relationship among three layers of governance: a strategic layer, a management layer, and an operational layer.

4.1 *Strategic Layer*

The strategic layer focuses on the interrelationships between the Assembly of Members (Europeum Assembly Board), Advisory Board, and strategic partners.

¹⁸The EDIC User Guide can be accessed at EC (2024): <https://digital-strategy.ec.europa.eu/en/policies/edic>.

¹⁹ ‘Section 3.5 of the User Guide’.

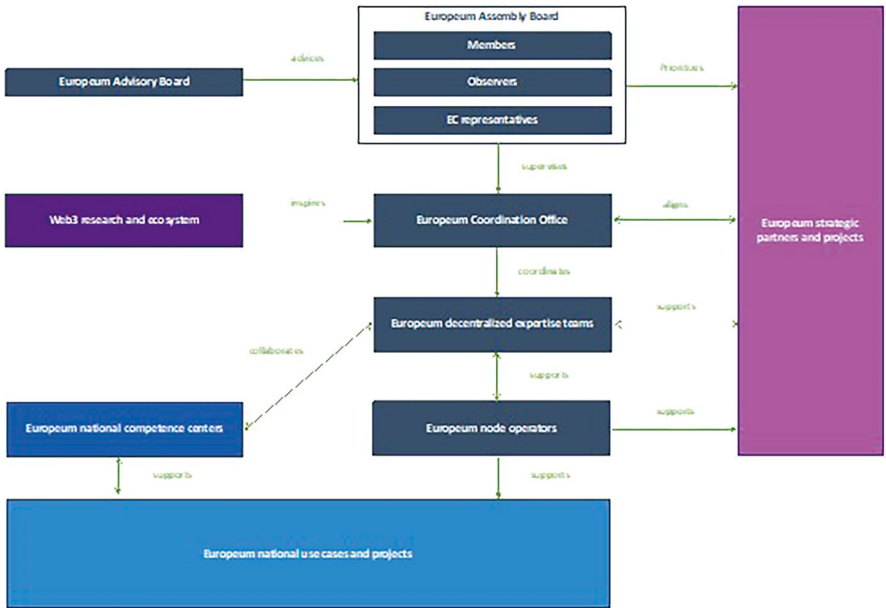


Fig. 1 Proposed governance formation for Europeum. Source: EDIC working group

The Assembly Board is set to serve as the decision-making body of Europeum, making decisions by a 2/3 majority vote. It comprises members, observers, and the EC representatives. Voting rights in the Assembly are reserved for contributing members. The Assembly Board not only acts as the decision-making body for Europeum’s policies but also serves a role similar to EBP, providing a platform for the EC and Member States (MS) to engage in EBSI governance. Its key responsibility is to supervise the activities of the Director and Strategic Orientation Board, which together form the coordination office of the governance structure.

The second component is the Europeum Advisory Board. Members are invited by the Director and appointed by the Assembly of Members, based on proposals from EU institutions and agencies. The Advisory Board’s main role is to provide advice to the Director and the Assembly of Members on specific applications of EBSI in support of EU policies. It also acts as the informal link between EU bodies and Europeum, ensuring policy alignment between the two organizations.

The third component includes representatives of strategic partners whose operations align closely with the responsibilities and tasks of EBSI. These partners may include representatives from other cross-national digital infrastructure projects and institutions or organizations focusing on standards and technologies supporting EBSI infrastructure.

4.2 *Management Layer*

The management layer of Europeum is primarily tasked with overseeing the organization's day-to-day operations, encompassing financial and administrative decisions, and serves as the central hub for coordinating its operational activities. This layer comprises the Director and the Europeum Coordination Office.

The Director, serving as the legal representative of Europeum, is entrusted with implementing decisions endorsed by the Assembly of Members. Assisted by the Coordination Office, the Director devises Europeum's implementation strategy, gathering input from various governance bodies. Moreover, the Director assumes responsibility for risk assessment, project coordination, ensuring compliance with established terms of reference by governance groups, organizing Assembly of Members meetings, and proposing updates and modifications to Europeum's regulations. Beyond administrative duties, the Director plays a pivotal role in coordinating diverse initiatives and components across the broader EBSI ecosystem and related projects.

The Europeum Coordination Office comprises the Director's staff and experts based in Brussels. Its primary function is to support the Assembly of Members and the Director in fulfilling their respective mandates. Serving as the central support hub, the Coordination Office aids in coordinating the activities of decentralized expertise teams and acts as the primary helpdesk for EBSI network operations.

4.3 *Operational Layer*

The operational layer constitutes a horizontal governance component within the EBSI ecosystem, facilitating the coordination of node operators, technical representatives, use case proponents, and various ad-hoc advisory and project groups. Currently, it operates as an extension of the existing ecosystem centered around EBSI, tasked with maintaining the technical infrastructure of EBSI and supporting user groups funded by various European and EBSI funding programs.

A notable development within the operational layer is the establishment of decentralized expertise teams. These teams will each specialize in distinct areas and engage closely with Europeum and its partners, clients, and stakeholders. The decentralized teams are envisioned to include: legal and governance, risk, and compliance (GRC) team; policy and business development team; communication and strategic account management team; architecture and specification team; security team; Dev and DevOps team; operational management team; support office team; operation security team; and technical development infra team.

While the legal responsibility of EBSI node operators falls under the jurisdiction of partner countries, their financial responsibilities are managed by Europeum. Additionally, efforts are underway to explore the establishment of national competence centres to support local actors in partner countries, potentially launching

between 2025–2026. These centres would serve as national support teams to aid in the uptake and development of new use cases at the local level.

Regarding financial structuring, the development and support of use cases are currently not considered part of the operational layer. Additional budgetary support for establishing and operating new use cases is expected to be provided through external contributions by use case partners from different EDIC members. Use case partners need not be Europeum members, nor do the use cases require the provision of a node. It is anticipated that if national competence centres are established, they may assume some tasks related to operational expenses currently overseen by Europeum.

5 Governance Challenges

In addition to the analysis of the working documents, we have conducted semi-structured interviews (a total of 7 interviews) to explore the key governance issues as perceived by the task force members in charge of the preparation of Europeum’s governance structure and formal statuses. The interviews were conducted in early 2024. At the time, the task force was composed of representatives of 8 founding members: Belgium, Croatia, Slovenia, Poland, Portugal, Romania, Italy, and Luxembourg. We interviewed one representative from each country case, except for Croatia.

The documents have been instrumental in identifying the key categories in the formation of the governance structure. However, we have used the semi-structured interviews to understand better the expectations from EDIC in the governance of EBSI, why certain governance choices are preferable over others and what the contingencies are.

During the interviews, we have asked the following questions:

1. What are the key governance challenges of EBSI? Top 3 governance challenges of the current governance structure of EBP.
2. Why EDIC is important for the governance of EBSI?
3. What are the specific governance challenges for EDIC in the short, mid, and long-term?

The expert interviews have provided deeper insights into the challenges and considerations surrounding the current governance model of EBSI. The respondents shed light on various issues related to decision-making structures, the transition from the EBP to Europeum, financial concerns, the preservation of institutional knowledge, and long-term strategic considerations. Below we highlight the key issues expressed by the respondents concerning the governance challenges of the EBP and the Europeum.

Governance challenges of EBP: All informants perceive the transition from EBP to Europeum as a positive evolution for the governance of EBSI. EBP is perceived as inadequate to manage day-to-day operations of EBSI in the existing model, and

the transition toward Europeum is expected to enhance the efficiency and scalability of the operations. Informants acknowledge that EBP provided a legitimate platform to facilitate idea exchange, but they advocate a need for a legal representative body aligned with EU laws for sustained EBSI success. Especially, a formal body capable of entering into legal agreements with purveyors and third parties is necessary to support the implementation of EBSI services in different country cases. For instance, if a node operator disappears overnight there are no legal obligations can be imposed on the service providers within the existing institutional framework.

Furthermore, through Europeum, MS' involvement and influence in shaping EBSI policies is expected to increase. In the Europeum governance, the EC's role in decision-making will diminish compared to the EBP, as their role will be reduced to a non-voting partner at the Assembly.

Another shortcoming of EBP is perceived in the effectiveness of the operational processes. Informants emphasize a formal bureaucratic structure would improve the effectiveness of the operations by accelerating decision-making processes and streamlining operations in comparison to EBP governance. They expect the institutional transition towards Europeum to enhance the formation of the EBSI services at an institutional and technical level and bring them faster to the level of large-scale implementation.

A third weakness of the EBP's governance structure relates to the concerns expressed by the informants about the political legitimacy of some country representatives taking part in the EBP policy group meetings. The country representative is expected to act as a channel between the political representatives of MS and the EBP. However, certain EBP representatives encounter challenges in persuading political counterparts of EBSI's significance, attributed to changes in cabinet structures and limited awareness of the EBSI blockchain among political representatives. These obstacles raise concerns about the legitimacy and representativeness of the EBP policy group in the governance of the EBSI ecosystem.

One last weakness highlighted by the task force pertains to the range of use cases initiated by the EBP. Initially, about a dozen use cases were selected across various domains, including education, notarization, supply chain management, and asylum management, to be promoted by the EBP in building their own network of participants and solutions. However, over time, only a few have progressed to a level of maturity suitable for support from EBSI services. This discontinuity in the chosen application areas is partly attributed to the lack of a central role within the existing governance structure of EBP capable of supporting and addressing the evolving needs of diverse ecosystems and stakeholders.

Governance challenges of Europeum: Despite a positive perception of transitioning to Europeum by the task force, convincing all EBP members to join Europeum proved difficult, with concerns raised about the absence of key member states like Germany and France. Related challenges include difficulties in persuading political representatives and potential changes in cabinet structures impacting decision-makers' understanding of EBSI.

The financial responsibilities of MS, doubts regarding EBSI's long-term financial viability, and uncertainties about its strategic position compared to alternative

systems add complexity to the transition. Respondents noted a more supportive environment for blockchain technologies during EBP's launch, without financial commitments attached to membership. Consequently, EBP created a more legitimate decision-making venue, unlike Europeum, which lacks key partners, potentially questioning its legitimacy. However, some interviewees view this as an opportunity for public representatives to justify the need for public infrastructure and financially support a new institution and supranational governance body. One respondent highlighted EBSI's importance as a European infrastructure guaranteeing sustainability and assurance for public services. For instance, a cross-border service backed by a European infrastructure should create trust and public value, offering an alternative to privately-owned or non-European initiatives.

Financial challenges emerged as a recurring theme in interviews. The allocation of budgets for Europeum operations in 2024 encountered obstacles, with some member states expressing doubts about EBSI's financial sustainability. Ambiguities surrounding EBSI's role and concerns about Europeum replacing EBP further complicate financial discussions. A broader agreement among member states and the EC to share funding responsibilities is recognized as necessary for EBSI's continued financial support. However, some respondents noted minimal expected financial commitments from Member States, highlighting a visionary incompatibility of blockchain with the perspectives of decision-makers in certain MS. The ongoing preparatory work on financial forecasts aims to achieve financial support from 12 MS by the end of 2024, gradually evolving to the support of all MS by 2026 when grants from existing DEP projects will end. To ensure long-term financial sustainability, various solutions are proposed, including expanding EBSI's use cases into new areas like copyright protection, integrating EBSI services with other technologies such as AI and IoT, adjusting to changes in financial instruments like the Digital Euro, and exploring private sector contributions.

Another challenge in the transition involves preserving institutional knowledge and capabilities within the current EBSI operational team. The potential reshuffling of key personnel within the EC and ongoing support for existing personnel during early Europeum operations create uncertainties. It is considered essential to have clear delineation of responsibilities, both financial and operational, between the EC and Europeum until certain capacities are established. Current plans suggest transferring the existing EBSI team to the newly established Europeum by the end of 2024. However, some interviewees express doubts about the EC's accelerated timeline to complete the transition of technical and operational governance responsibilities to Europeum without securing necessary funds and support beforehand.

While the immediate governance challenges primarily revolve around political, administrative, and financial issues, mid-term concerns encompass interoperability challenges, such as integrating with additional blockchains, ensuring compatibility with cybersecurity standards across different jurisdictions, and broadening the scope of use cases. Some stakeholders have also voiced apprehensions about certain countries' reluctance to develop cross-border services, which could potentially hinder the advancement of EBSI use cases. Moreover, discussions on long-term challenges have been somewhat limited within the task force, but emerging technological

issues related to Web3 and the anticipated rise of AI agents assuming prominent roles in the public sphere underscore the need for a robust and decentralized trust model. Consequently, the sustainability of the EBSI blockchain is perceived to be intricately linked to the digital transformation of the European public sphere. Finally, some respondents envision a decentralized governance structure for Europeum, proposing an exploration of Decentralized Autonomous Organization (DAO) as a prospective governance model for EBSI. Through this model, a more democratic governance framework allowing citizens to directly influence public policies may be realized.

6 Discussion

The transition from the EBP to Europeum represents a strategic response to the inherent limitations of the collaborative governance structure that characterized the EBP. While the EBP primarily served as a discussion forum, decision-making authority regarding the evolution of EBSI largely resided with the EC. Consequently, the pace and funding of the EBSI project were determined by institutional conditions within the EC, resulting in a lack of shared responsibility among its founding partners. This gap in collaborative governance led to limited public visibility and stakeholder engagement in the EBSI project since its inception in 2018. Furthermore, the informal role of the EBP, which was unable to enter contractual relationships with service providers such as node operators and wallet providers, hindered the establishment of a legitimate trust framework to support the implementation of cross-border use cases.

Hence, the transition to Europeum represents an effort to address the shortcomings of the EBP in fully operationalizing EBSI to support services in cross-border settings. However, the regulatory and operational needs alone do not entirely explain why the EC and EBP members were willing to establish a governance structure reminiscent of bureaucratic organizations within Union bodies rather than exploring alternative governance forms such as DAOs, as suggested by some experts during the interviews as a potential future direction. The literature on blockchain governance argues that blockchain, as a decentralized technology, can unlock new forms of governance that can help network entities self-govern themselves more efficiently and effectively.²⁰ Given the established capabilities of the EBSI network, it is important to reflect on the possible theoretical explanations that may have led the EC and the task force to pursue the governance structure as elaborated earlier.

As a starting point, isomorphism theories can provide insight into the decisions driving the institutionalization of Europeum. Isomorphism theories suggest that when organizations face uncertainty regarding organizational change, they adhere

²⁰ See De Filippi and Wright (2018), Miscione et al. (2019), Werbach (2020), Tan et al. (2022), Tan and Rodriguez (2023).

to a logic of appropriateness within the constraints imposed by the institutional environment. This adherence often leads to policy convergence and institutional homogenization. Isomorphism theories posit that such change can be instigated by coercive, mimetic, and normative processes.²¹

Coercive isomorphism suggests that organizations respond to external pressures within their institutional environment to gain legitimacy. Such external pressures may be reflected in the formation of the member state-led decision-making structure of Europeum.

During the development of EBSI, the EC had direct control over operations and decisions regarding use cases and technological choices, with operational expenses covered by EU resources through new or established funding streams. The leading role of the EC was instrumental in establishing an operational infrastructure to support the EBSI infrastructure. However, in the directive concerning the establishment of the EDICs, the role of the EC is reduced to a more advisory role in governance where the actual decision-making power is supposed to be held by the MS except for activities directly funded by EU programmes.²² There seem to be two coercive pressures in the preference towards a reduced role for the EC representatives.

Firstly, in EDIC governance, MS are expected to bear the financial burden of operational processes, and understandably, MS should have a larger say in EBSI governance decision-making. However, this does not justify alone why EC representative roles are proposed without any voting rights in the Assembly. One explanation could be that the EU does not have a single market for digital networks and services, but 27 national markets with different supply and demand conditions, network architectures, different levels of coverage of very high-capacity networks, different national spectrum authorization procedures, conditions, and timing, as well as different regulatory approaches.²³ Although secure and sustainable digital infrastructures are mentioned as one of the four fundamental objectives of the EU's Digital Decade Policy Programme 2030, the lack of a single market for digital infrastructures prevents the EC from taking over decision-making functions on behalf of the MS.

Secondly, the transition towards a comprehensive implementation environment for verifying cross-border public services in higher education, identity management, and social security domains,²⁴ with a governance model dominated by the EC in decision-making authority, risks encroaching on policy areas traditionally perceived as the domain of MS. Given the noted bureaucratic resistance in certain MS towards cross-border digital public services, as highlighted by experts, increased involvement of MS in decision-making may be necessary to foster legitimacy in Europeum's governance, particularly in light of power dynamics within European governance. This aligns well with coercive isomorphism mechanisms, emphasizing

²¹ See DiMaggio and Powell (1983), DiMaggio and Powell (1991).

²² Art. 16(2) Decision 2022/2481.

²³ EC (2024b).

²⁴ See Tan et al. (2023).

formal and informal pressures exerted by organizations within the institutional cultural context.²⁵

The second form of isomorphism, mimetic isomorphism, postulates that in an uncertain situation, an organization looks for best practices outside of its organizational space. Mimetic isomorphism can elucidate why the Assembly and Director are established as the two mandatory governance units in the proposed EDICs. The Decision²⁶ obliges EDIC applicants to establish an Assembly of Members as the decision-making body, and the appointment of a Director by the Assembly as the executive body and legal representative of the EDIC mirrors the institutional structure of the EU, with a Directorate analogous to the European Commission and an Assembly akin to the Council of Ministers. Our interpretation is that through enforcing such an organizational design, the EC seeks to establish legitimacy for EDICs within the EU institutions familiar with this supranational governance formation.

The third form of normative isomorphism may explain Europeum's departure from the EBP's collaborative model towards a more bureaucratic structure. Normative isomorphism suggests that professional networks spanning organizations often reflect the educational and professional background of the network actors. Task force members and EC officials, predominantly civil servants, perceive a director-led organization, mirroring established bureaucratic norms with clear operational parts that comply with a hierarchical and bureaucratic decision-making structure to be more effective in bringing the EBSI ecosystem into a full implementation environment faster and more efficiently. Even though this transition will require the MS to take additional financial and operational responsibilities in sustaining this digital infrastructure and the operational costs of a new organization, they expect a hierarchical organization with a clear legally responsible entity to be more suitable for the operational needs.

Nonetheless, isomorphism theories do not fully elucidate the organizational decisions made in forming Europeum's governance structure, especially regarding the governance of the operational layer. In the proposed structure (see Fig. 1), a modular approach is adopted wherein decentralized expert teams and node operators supporting the EBSI infrastructure are organized into autonomous subgroups containing executive, advisory, and technical components. This preference can be explained by the *mirroring thesis* proposed by Baldwin and Clark,²⁷ which suggests that technical dependencies shape organizational ties because actors performing interdependent tasks necessitate organizational ties such as collocation, communication linkage, and common dispute resolution mechanisms. This thesis also aligns with Conway's Law,²⁸ which suggests that organizations designing systems tend to produce designs reflective of their communication structures. The task force is convened by representatives from each founding member, and the

²⁵ See DiMaggio and Powell (1991).

²⁶ Art. 16(1) Decision 2022/2481.

²⁷ See Baldwin and Clark (2000).

²⁸ See Conway (1968).

decentralization of various operational tasks across these members aligns well with the nature of the task force. Moreover, the operational framework of the EBP has already established node operators in various country cases, while the funding mechanisms of the Digital Europe Programme have assembled project teams for distinct use cases centered around specific topics, each with its internal communication and coordination mechanisms. By maintaining the separation between node operators and use cases as distinct operational modules, the proposed governance model aims to safeguard knowledge and existing collaborations during the transition of capabilities into the new Europeum organization.

In summary, the transition from EBP to Europeum represents a strategic endeavor to establish legitimacy by creating a recognizable multinational organizational structure within European governance. In this sense, Europeum's governance formation seems to be influenced more by institutional conditions in European governance than by the inherent technological properties of blockchain systems.²⁹ However, the decision to maintain modular structures for use cases, decentralized entities, and node operators, granting them autonomy in internal operations, aligns with the technical requirements of running a blockchain infrastructure. Therefore, we observe that until a critical mass of users is achieved, where both public and private actors actively contribute to the ecosystem as users and developers, and regulatory frameworks provide conditions that support alternative governance forms such as DAOs, traditional organizational and institutional structures remain crucial for garnering support and contributions from MS and EU institutions.

7 Conclusion

The emergence of Europeum represents a significant milestone in the governance of blockchain-based infrastructures at the supranational level, marking the establishment of an international organization for the governance of a state-led blockchain infrastructure. EDICs, within the framework of the Digital Decade Programme, offer a novel institutional response aimed at addressing the multifaceted challenges inherent in cross-border digital projects.

The establishment of Europeum reflects a strategic response to the limitations of the previous collaborative governance structure, the European Blockchain Partnership (EBP). While the EBP provided a forum for discussion, decision-making authority primarily rested with the European Commission (EC), resulting in a lack of shared responsibility among participating Member States (MS) and limited stakeholder engagement. The transition to Europeum underscores the necessity for a more robust governance framework capable of addressing the evolving needs of blockchain-based initiatives. Insights from expert interviews emphasize the importance of addressing immediate transition challenges, financial concerns, and

²⁹ See Tan et al. (2022).

providing a clear vision for long-term sustainability and growth. A robust and adaptable governance structure for EBSI, coupled with effective communication and collaboration among member states, emerges as vital for the success of this European blockchain initiative.

Isomorphism theories shed light on the organizational choices driving the formation of EDICs. Coercive isomorphism highlights external pressures shaping Europeum's governance structures, including the shift towards a full implementation environment for cross-border public services. Mimetic isomorphism elucidates the replication of familiar organizational structures from established EU institutions, enhancing legitimacy within the European governance framework. Normative isomorphism underscores the importance of hierarchical and bureaucratic structures in facilitating effective decision-making and operational efficiency.

Furthermore, Europeum's modular approach reflects a strategic alignment with technical requirements inherent in running blockchain infrastructures. Granting autonomy to decentralized entities and node operators aims to foster innovation while preserving existing collaborations and knowledge transfer from the EBP era.

In essence, EDICs represent a significant evolution in the governance of digital infrastructures, offering a pathway towards enhanced collaboration, innovation, and resilience in the European digital landscape. As Europeum and future EDICs develop, their role as facilitators of digital transformation and cross-border cooperation will be crucial in shaping the digital future of the European Union. Looking forward, the progression of Europeum holds potential for a more decentralized and adaptable governance model, which may explore alternative forms such as Decentralized Autonomous Organizations (DAOs). As regulatory frameworks evolve and technological advancements progress, Europeum is positioned to navigate intricate landscapes, promote stakeholder engagement, and foster innovation in blockchain-based infrastructures. Pursuing a decentralized system that empowers user groups in various sectoral domains may not only be an option but also a necessity in the future.

References

- Baldwin CY, Clark KB (2000) Design rules, volume 1: the power of modularity. The MIT Press
- Benchaya Gans R, Ubacht J, Janssen M (2022) Governance and societal impact of blockchain-based self-sovereign identities. *Polic Soc* 41(3):402–413
- Cagigas D, Clifton J, Diaz-Fuentes D, Fernandez-Gutierrez M (2021) Blockchain for public services: a systematic literature review. *IEEE Access* 9:13904–13921
- Conway M (1968) How do committees invent? *Datamation* 14(4):28–31
- de Filippi P, Wright A (2018) Blockchain and the law: the rule of code. Harvard University Press
- DiMaggio PJ, Powell WW (1983) The iron cage revisited: institutional isomorphism and collective rationality in organizational fields. *Am Sociol Rev* 48(2):147–160
- DiMaggio PJ, Powell WW (1991) Introduction. In: Powell WW, DiMaggio PJ (eds) *The new institutionalism in organizational analysis*. University of Chicago Press, pp 1–38
- European Blockchain Services Infrastructure (EBSI) (2024a) EBSI hub. European Blockchain Services Infrastructure (EBSI) <https://hub.ebsi.eu/>

- European Blockchain Services Infrastructure (EBSI) (2024b) Home. European Blockchain Services Infrastructure (EBSI) (2024c) Success stories. European Blockchain Services Infrastructure (EBSI). <https://ec.europa.eu/digital-building-qblocks/sites/display/EBSI/Verifiable+Credentials+Success+Stories>
- European Commission (2022) European blockchain partnership. European Commission. <https://digital-strategy.ec.europa.eu/en/policies/blockchain-partnership>
- European Commission (2024) European digital infrastructure consortium (EDIC). European Commission. <https://digital-strategy.ec.europa.eu/en/policies/edic>
- Miscione G, Goerke T, Klein S, Schwabe G, Ziolkowski R (2019) Hanseatic governance: understanding Blockchain as organizational technology. In: Fortieth International Conference on Information Systems, Munich 2019, 1–17
- Services Infrastructure (EBSI). <https://ec.europa.eu/digital-building-blocks/sites/display/EBSI/Home>
- Tan E, Rodriguez AP (2023) Paths to citizens-controlled coproduction: the use of Blockchain technology in digital coproduction. *Public Manag Rev*:1–20
- Tan E, Mahula S, Crompvoets J (2022) Blockchain governance in the public sector: a conceptual framework for public management. *Gov Inf Q* 39(101625):1–11
- Tan E, Lerouge E, Du Caju J, Du Seuil D (2023) Verification of education credentials on European Blockchain Services Infrastructure (EBSI): action research in a cross-border use case between Belgium and Italy. *Big Data Cogn Comput* 7(79):2–20
- Werbach K (2020) The Siren song: algorithmic governance by Blockchain. In: Werbach K (ed) *After the digital tornado: networks, algorithms, humanity*. Cambridge University Press, pp 215–240

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.



Privacy-By-Design in the Belgian Public Sector



Pseudonymising and Joining Personal Data Fragmented Over Multiple Organizations

Kristof Verslype and Bart De Decker

Abstract In various countries and regions, sensitive citizen data —encompassing social, medical, and fiscal information— is typically dispersed among government authorities and private entities. This fragmentation poses a challenge to unlocking the full potential of combined datasets, which could yield invaluable insights across sectors such as healthcare, fraud detection, and evidence-based policymaking.

To safeguard privacy, data integration —i.e., joining the data— occurs on a project-specific basis rather than on a massive scale. Each request for access to fragmented personal data initiates a separate project, where only the minimal necessary data is combined and pseudonymised. Pseudonymisation involves converting national citizen identifiers into unique codes relevant only at the project level. While today's implementation strongly relies on trusted third parties, this approach is sub-optimal due to inherent security risks and the potential for projects to become cumbersome, slow, expensive, and labour-intensive.

This chapter delves into and compares two innovative solutions which exemplify privacy-by-design in the public sector, leading to significant efficiency gains in time and resources. Notably, the reliance on third party trust is minimised. These advancements offer new possibilities for more intricate research, yielding insights that positively impact healthcare, policymaking, competitiveness, and society at large.

1 Introduction

This section clarifies how technology can improve privacy, introduces the use case of unlocking insights in dispersed data and outlines our contributions.

K. Verslype (✉)
Smals Research, Brussels, Belgium
e-mail: kristof.verslype@smals.be

B. De Decker
imec-Distrinet, KU Leuven, Leuven, Belgium
e-mail: bart.dedecker@cs.kuleuven.be

1.1 *Privacy-Enhancing Technologies*

The privacy of citizens is universally acknowledged as a fundamental right, although specific definitions may vary across cultures.¹ A growing segment of the global population benefits from privacy laws, with e.g., the GDPR being enforced in the European Union since 2018,² the United Kingdom adopting the UK GDPR in 2020, and California implementing the CCPA since 2018. Numerous other states within the United States either have existing privacy legislation or are actively working on its development.³

The processing of Personally Identifying Information (PII), i.e., information linked to an identifiable person, has become more accessible than ever. This leads to numerous potential benefits for individuals, businesses, public actors, and society as a whole. However, as a society, we must seek ways to achieve this while sufficiently respecting the privacy of the individuals involved.

Traditionally, privacy and data utility are viewed as competing interests —increasing one often comes at the expense of the other.⁴ However, ongoing research is dedicated to developing *Privacy-Enhancing Technologies (PETs)* to bridge this gap while minimising reliance on trusted third parties.⁵ Examples of PETs include attribute-based credentials,⁶ multi-party computation,⁷ differential privacy,⁸ homomorphic encryption,⁹ and private set intersection.¹⁰

PETs serve as foundational elements for *privacy-by-design*, according to which privacy is integrated as a fundamental requirement throughout the entire design and development process of systems.¹¹ This approach aims to reconcile the seemingly conflicting interests of data utility and citizens' privacy. Presently, the adoption of PETs is on the rise, albeit limited. Implementing current generations of PETs can be challenging, significantly impacting system complexity, performance, flexibility, and maintainability. Tailored PETs are necessary to meet specific requirements essential for balancing data use and privacy.¹²

¹ See Rubinfeld (1989).

² See GDPR (2016).

³ See Bellamy and Fernandez (2023).

⁴ See Wieringa et al. (2021).

⁵ See Kaaniche et al. (2020).

⁶ See Rannenberg et al. (2015).

⁷ See Zhao et al. (2019).

⁸ See Li et al. (2021).

⁹ See Marcolla et al. (2022).

¹⁰ See Morales et al. (2023).

¹¹ See Cavoukian (2009).

¹² See OECD (2023).

1.2 *Unlocking Insights: Joining Fragmented Personal Data*

High-quality data plays a pivotal role in advancing cutting-edge usage but particularly also enabling research across both public and private sectors. However, this valuable resource of data is often fragmented, scattered among various government agencies and commercial entities. While individual datasets are interesting in their own right, added value lies in combining them, potentially enhancing the accuracy of the analysis and even reducing costs.

The merging of personal data scattered across multiple organizations can occur through structural or on-demand exchanges. Structural exchanges support long-term monitoring, such as research on morbidity.¹³ On-demand exchanges, on the other hand, are often shorter-term and result from specific requests by researchers, whether in academia, the public sector, or the private sector. For instance, academic researchers recently investigated whether the use of certain medications for Multiple Sclerosis (MS) correlated with an increased cancer risk.¹⁴

To protect privacy, data combination is generally not conducted on a massive scale. Instead, each request for access to fragmented personal data typically initiates a separate project, where only the minimal necessary personal data is combined and pseudonymised.¹⁵ (Single identifier) pseudonymisation involves converting national citizen identifiers into distinct codes relevant solely at the level of the specific project. This approach reduces re-identification risk while remaining suitable for data analysis and other processing tasks.¹⁶

Efforts to safeguard citizens' privacy have, for instance, led to proposals for privacy preserving distributed data mining, relying on secure multi-party computations.¹⁷ However, such approaches face limitations in practice due to architectural and computational complexity. Consequently, organizations often revert to basic techniques, demanding high trust in multiple parties. Unfortunately, these challenges hinder the full realization of data's potential.

1.3 *Our Contribution*

This chapter delves into and compares two innovative PETs developed by the Research & Innovation unit of Smals, an IT provider serving the Belgian public sector. Executed in close collaboration with KU Leuven, these solutions exemplify privacy-by-design principles in the public sector, resulting in significant efficiency

¹³ See eHealth Beraadslaging 19/222 (2019).

¹⁴ See eHealth Beraadslaging 20/020 (2020).

¹⁵ See Neubauer and Heurix (2011).

¹⁶ See Jensen et al. (2019).

¹⁷ See Archer et al. (2018).

gains in time and resources and a substantial reduction in the reliance on third-party trust.

In both instances, the resulting pseudonymised data undergoes further processing within a secure environment tailored for privacy-preserving data mining. Despite pseudonymisation, potential identification risks persist due to thorough analysis based on *quasi-identifiers*—combinations of attribute values that are often unique.¹⁸

The proposed solutions pave the way for additional and more intricate research, yielding insights that could contribute to advancements in healthcare, policymaking, competitiveness, and society at large.

Section 2 and 3 introduce current practices and preliminaries respectively. Section 4 continues with two tangible projects necessitating the joining and pseudonymisation of fragmented data. The first project exemplifies the initial solution, leveraging an existing service on the Belgian eHealth platform, which is elaborated in Sect. 5. The second project serves as an illustration for the enhanced solution, which is discussed in detail in Sect. 6. This more sophisticated approach proves beneficial in specific cases where the initial, more generic approach falls short. Section 7 discusses on how the presented solutions align and comply with GDPR. Section 8 concludes this chapter.

2 Current Practices

For privacy reasons, existing solutions strongly rely on trusted parties and must comply with legal regulations to control the collection, linking and pseudonymisation of data before it can be used for analysis. Examples of such trusted parties trusted include the *Census Bureau* in the US, the *Research Data Centers (RDCs)* in Germany, the *National Institute of Statistics and Economic Studies (INSEE)* in France and the *Federal Statistical Office* in Switzerland.¹⁹ These parties receive sensitive data, including personal identifiers, which allow for linking datasets from different providers. The data either remains in a strongly secured central data warehouse for local processing, or is joined, pseudonymised, and provided to another entity. In either case, the trusted party needs a high level of trust to properly protect and manage the collected data.

Unwanted data leakage may occur when a subset of personal identifiers is used to request data from individual data sources. For example, the presence of a citizen in the result set of one data source—e.g., citizens with Multiple Sclerosis—could be used to request additional data about this citizen from another data source, leaking information about the citizen's presence in that subset. As data joining and

¹⁸ See de Montjoye et al. (2013); de Montjoye et al. (2015); Sweeney (2000).

¹⁹ See Swiss Academy for Humanities and Social Sciences (SAGW) and Swiss Academies of arts and sciences (2020).

pseudonymising data for research purposes are frequent, these data leakages can become extensive.

In summary, existing approaches have significant shortcomings. A distributed, low-trust, privacy-preserving, efficient (in terms of computation and communication), and practical approach is necessary for analysing personal data dispersed across multiple data sources.

3 Preliminaries

This section introduces the main concepts related to cryptography and pseudonymisation.

3.1 Cryptography

Encryption involves encoding information in a manner that restricts access solely to authorized parties. Algorithms are employed to convert data into an unreadable form ('encrypt'), known as *ciphertext*. Only by using a key can this ciphertext be transformed back ('decrypt') into its original, readable form, known as *plaintext*. Encryption is generally *probabilistic*; encrypting the same plaintext twice results in different ciphertexts.

A *key* is a large number which should be unfeasible to guess. In *symmetric cryptography*, both encryption and decryption occur using the same *secret key*. In contrast, *public-key cryptography* involves encryption with a publicly known *public key*, while decryption necessitates a *private key* held by a single entity.

3.2 Pseudonymisation

A *citizen identifier (global identifier)* is a unique number attributed by the government to a citizen or resident for unambiguous identification (e.g., social security number).

A *pseudonym (local identifier)* is a unique code that replaces the citizen identifier within a specifically defined scope (e.g., fertility data). It does not directly identify the individual.

Pseudonymisation involves replacing citizen identifiers by pseudonyms. The reverse operation is *identification*. Legal pseudonymisation is more stringent (see Sect. 7).

We define *pseudojoining* as joining and pseudonymizing personal data originating from multiple data sources.

The *collector* is positioned between the *data sources* and the *researcher*; it assembles the pseudojoined data based on input from various data sources and authorizes the researcher to perform controlled analyses on this data, without disclosing the sensitive data itself. While the collector only receives the minimally required pseudonymized data, trust in its proper handling remains imperative.

The *translator* pseudonymises citizen identifiers and re-identifies the citizens from these pseudonyms.

4 Concrete Projects in the Belgian Public Sector

Pseudojoining is a common practice in data integration. This section delves into two specific cases. While in the first case, all data sources are able to *autonomously* select the records relevant for the pseudojoin project, this is impossible in the second case. Descriptions and legal foundations of these and other pseudojoin projects in the Belgian healthcare and social security are publicly available.²⁰

4.1 Case 1: Epidemiology of Diabetes Patients

The aggregation of medical data from various hospitals, aimed at understanding the epidemiology²¹ of patients with diabetes is a critical endeavour conducted by *Sciensano*, an esteemed research institute, and the national public health institute of Belgium. Legally grounded in Deliberation 13/093 of October 22, 2013, by the Social Security and Health Chamber of the Information Security Committee,²² this initiative involves a carefully orchestrated data flow described in the deliberation. Figure 1 shows the high-level process for one of the many involved citizens, known under identifier *id*.

1. Each hospital with relevant data about this citizen randomly generates a unique number; the *ephemeral pseudonym* (nym_{id}^1 resp. nym_{id}^2) and transmits the citizen's data ($data_{id}^1$ resp. $data_{id}^2$) to Sciensano together with this ephemeral pseudonym.
2. Each hospital with relevant data encrypts this ephemeral pseudonym with Sciensano's public key (resulting in enc_{id}^1 respectively enc_{id}^2). Only Sciensano possesses the capability to decrypt these ciphertexts. The result is sent to the translator *Batch Codage*,²³ together with the citizen identifier (*id*).

²⁰ Kruispuntbank Sociale Zekerheid (n.d.); eHealth Beraadslagingen (n.d.).

²¹ Epidemiology involves the study and analysis of the distribution (who, when, and where), patterns, and determinants of health and disease conditions in a defined population.

²² eHealth Beraadslaging 13/093 (2013).

²³ eHealth Pseudonimiseren (n.d.).

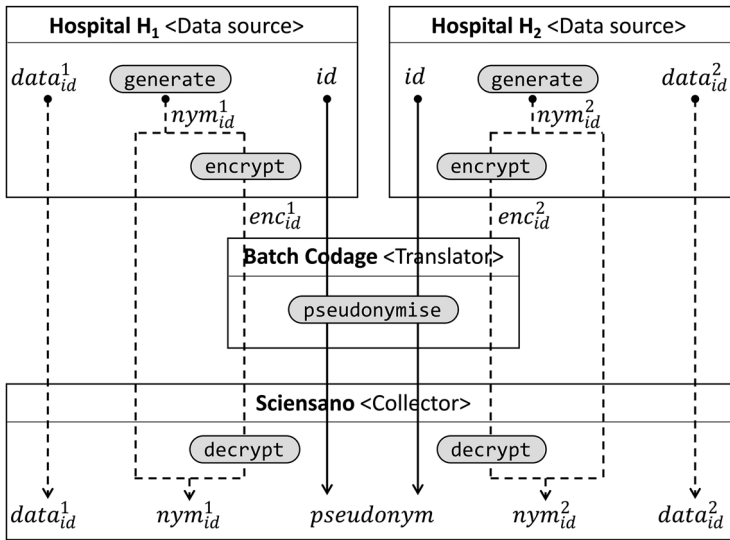


Fig. 1 Pseudojoining data from hospitals to improve epidemiology of diabetes patients

3. Upon receipt, Batch Codage translates (encrypts) this citizen identifier into a project-specific pseudonym (*pseudonym*) and delivers this pseudonym, along with the encrypted ephemeral pseudonym, to Sciensano; it sends $(pseudonym, enc_{id}^1)$ to the collector after receipt of (id, enc_{id}^1) and $(pseudonym, enc_{id}^2)$ after receipt of (id, enc_{id}^2) .
4. With its private key, Sciensano decrypts the received ciphertexts (enc_{id}^1 and enc_{id}^2) into ephemeral pseudonyms (nym_{id}^1 and nym_{id}^2), enabling it to link (hospital-specific) ephemeral pseudonyms to project-specific pseudonyms. Since it already knew the link between data and ephemeral pseudonyms, Sciensano can now link data about the same citizen, but originating from different hospitals to the project-specific pseudonym, and, hence, to each other.

While this approach enables Sciensano to merge hospital data without knowledge of the national citizen identifier, we identify a drawback: sensitive data leaks to the Batch Codage. This service gains insight into which citizens are involved in the project, thereby discerning individuals with diabetes. Furthermore, it learns which hospitals possess diabetes-related information about each citizen. Section 5 elaborates on how the existing infrastructure can be leveraged to address this data leakage.

4.2 Case 2: Cancer Risk for Patients with Multiple Sclerosis

Researchers aimed to address the question:

Do patients with Multiple Sclerosis (MS) who take medications containing teriflunomide or alemtuzumab face an increased cancer risk compared to MS patients treated with other medications?

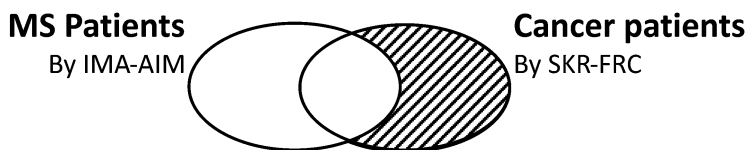


Fig. 2 Set of citizens with MS and set of citizens with cancer

The legal basis for this pseudjoin project is outlined in Deliberation No. 20/020 (14 January 2020) of the Social Security and Health Chamber of the Information Security Committee.²⁴

To achieve this goal, data from two distinct sources must be amalgamated:

- The *IMA-AIM* (*Intermutualist Agency*) possesses information on citizens with MS, including their medication regimens.
- The *SKR-FRC* (*Cancer Register Foundation*) maintains data on individuals diagnosed with cancer, including details about the type and timing of the diagnoses.

The Venn diagram in Fig. 2 depicts the intersection of two sets: MS patients and cancer patients. Researchers should only be granted access to pseudonymised records of citizens belonging to the unshaded regions.

However, a challenge arises: the collector needs to obtain cancer data only about citizens with MS, while the SKR-FRC should not learn the identities of those diagnosed with MS. The question is:

How can SKR-FRC transmit solely records of MS patients to the collector without learning who has MS?

Unfortunately, existing solutions heavily rely on trusted intermediaries, are custom-made, cumbersome, slow, expensive, labour-intensive, and fraught with security and privacy risks. Therefore, Sect. 6 introduces a *distributed* solution, eliminating the need to rely on a trusted translator; translations happen collectively.

5 eHealth Translator

Since December 2023, the Belgian eHealth platform has introduced a high security translator.²⁵ This service plays a pivotal role in securing various aspects of health-care data, including electronic prescriptions, medication regimens, medical record summaries, and information related to vaccinations, allergies and intolerances, and fertility.

²⁴ eHealth Beraadslaging 20/020 (2020).

²⁵ eHealth Pseudonimisering (n.d.); Verslype (2023).

Section 5.1 introduces the Belgian eHealth platform. Sections 5.2 and 5.3 delve into the roles, operations, and interactions involved in the pseudonymisation process. Section 5.4 meticulously examines the security properties associated with pseudonymisation through the eHealth translator. Following this, Sect. 5.5 elucidates the application of this service in *pseudojoin* projects.

5.1 The Belgian eHealth Platform

The *Belgian eHealth platform* is a prominent public social security institution in Belgium.²⁶ It provides an electronic platform that facilitates secure and efficient information access for all stakeholders in public health, including care providers, institutions, health insurance funds, and patients.

The primary objective of the platform is to enhance electronic services and information exchange among various entities within the Belgian healthcare landscape. It plays a crucial role in modernizing healthcare services, promoting efficiency, and safeguarding sensitive health-related data.

5.2 Roles & Operations

In scenarios without pseudonymisation, two primary roles come into play: the *owner*, responsible for managing medical personal data, and the *client*, who sends requests to the owner. Three illustrative scenarios highlight these roles and their interactions:

- **Scenario 1:** A Doctor (client) requests the prescription service (owner) to register an electronic prescription.
- **Scenario 2:** A pharmacist (client) requests a prescription for a specific citizen from the prescription service (owner).
- **Scenario 3:** A doctor (client) asks the prescription service (owner) to view the electronic prescriptions issued to a specific patient.

In scenarios with pseudonymisation, services store personal data under pseudonyms instead of citizen identifiers. It requires an extra role; the *translator*, responsible for translating the patient's identifier into a pseudonym (*pseudonymise* operation) and vice versa, converting a pseudonym back into a citizen identifier (*identify* operation). Additionally, the doctor's identification number can undergo pseudonymisation.

²⁶ See eHealth Portaal ([n.d.](#)).

Seamless communication between various services on the eHealth platform is essential. *TherLink*,²⁷ for example, is a service contacted by many other eHealth services for authorization purposes. It manages the therapeutic relationships between health care providers and patients, which enables it to answer questions such as “*Is this doctor the current general practitioner of that citizen?*”

To minimize identification risks, a citizen is known under a different pseudonym per service. Therefore, inter-service communication requires a translator able to convert one service’s pseudonym (owner) into another service’s pseudonym (owner) without revealing the underlying citizen identifier (*convert* operation).

In summary, the translator plays a pivotal role by supporting three crucial operations: *pseudonymise*, *identify*, and *convert*.

5.3 Interactions

The introduction of the translator significantly impacts the information flow. Based on scenario 1, Figs. 3 and 4 illustrate two distinct modes: *relay mode* positions the translator between the client and owner, while *reply mode* positions it more discreetly alongside these entities.

In *relay mode* the translator acts as a conduit. When faced with a pseudonymise scenario, it receives identifiers associated with —potentially personal— data, applies the pseudonymise operation on these identifiers, and then forwards the data under the pseudonyms. This mode is adopted by the eHealth service *Batch Codage*,²⁸ an established translator used by Sciensano,²⁹ as well as by the privacy-preserving scheme proposed by Verheul.³⁰

In *reply mode*, the translator receives and processes requests for pseudonymisation, identification, or conversion and returns the result to the requesting party. This mode is adopted, among others, by *eHealth WS SEALS* and *ScrambleDB*.³¹

Both modes offer distinct advantages and challenges. However, reply mode aligns more effectively with the requirements of the Belgian eHealth platform:

- **No man-in-the-middle:** In *reply mode*, the translator deals exclusively with pseudonyms and identifiers, avoiding any —potentially large— associated data. This approach prevents burdening the translator with extensive data handling, whether encrypted or not, and may cause extra delays.

²⁷ See eHealth TherLink (n.d.).

²⁸ See eHealth Pseudonimiser (n.d.).

²⁹ See Sciensano (n.d.).

³⁰ See Verheul et al. (2016).

³¹ See eHealth Pseudonimiser (n.d.) and Lehmann (2019).

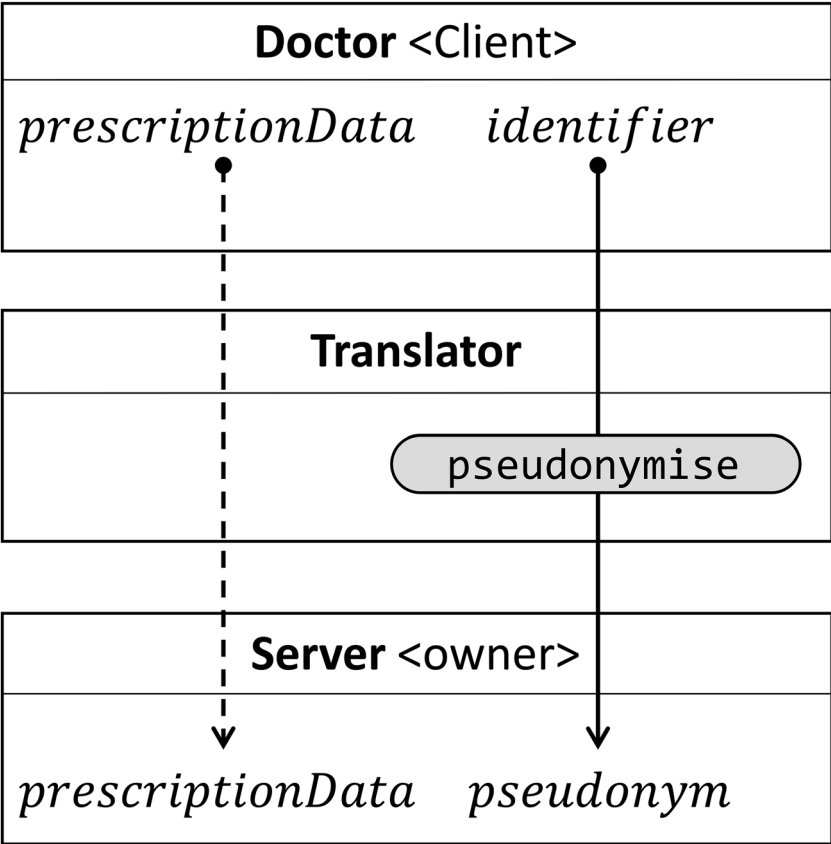


Fig. 3 Scenario 1 in relay mode

- Mass data forwarding in *relay mode* can be avoided at the expense of increased communication complexity. Managing two flows—one with an intermediary translator and one direct—becomes unavoidable (see Fig. 1).
- **Low intrusiveness:** By using a single, direct communication channel between client and owner (or owner and owner), reply mode minimizes disruption of interactions and brings low-level interactions closer to business interactions.

5.4 Security Properties

Given the sensitivity of medical data, ensuring robust security is paramount. Using an external translator realises *separation of duties*; a fundamental security principle whereby more than one entity is required to complete a task.

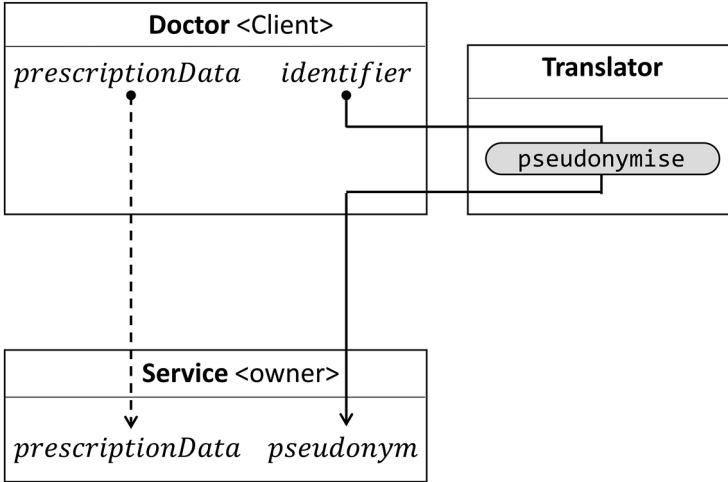


Fig. 4 Scenario 1 in reply mode

Additionally, the processes in and of themselves to pseudonymise, identify and convert must adhere to rigorous security standards. This mandates that none of the involved parties learns more information than strictly necessary to fulfil their duties: the client (e.g., a doctor) should not gain access to pseudonyms, the owner (service) should remain unaware of citizen identifiers, and the translator should remain oblivious to both. To accomplish this, we extend the basic flow outlined in Fig. 4 into the high-security flow depicted in Fig. 5, encompassing four key steps:

1. **Blinding by client:** The client encrypts the citizen identifier using a freshly generated ephemeral key —a process known as blinding. The blinded identifier is then sent to the translator, effectively concealing the citizen identifier from the translator.
2. **Translation by translator:** The translator converts the blinded identifier into a blinded pseudonym, encrypting it along with context information, such as the current time. This encrypted blinded pseudonym is returned to the client.
3. **Unblinding by client:** The client removes the blinding and forwards the resulting encrypted pseudonym, along with the prescription data, to the server. Only the client can unblind since no one else knows the ephemeral key. (The capability to unblind encrypted blinded values is a special property.)
4. **Decryption by owner:** The owner decrypts the encrypted pseudonym, validates the context information (e.g., ensuring the timestamp is not older than a specified duration), and stores the received prescription data under the resulting pseudonym. The verification of context data prevents unauthorized reuse of the ciphertext.

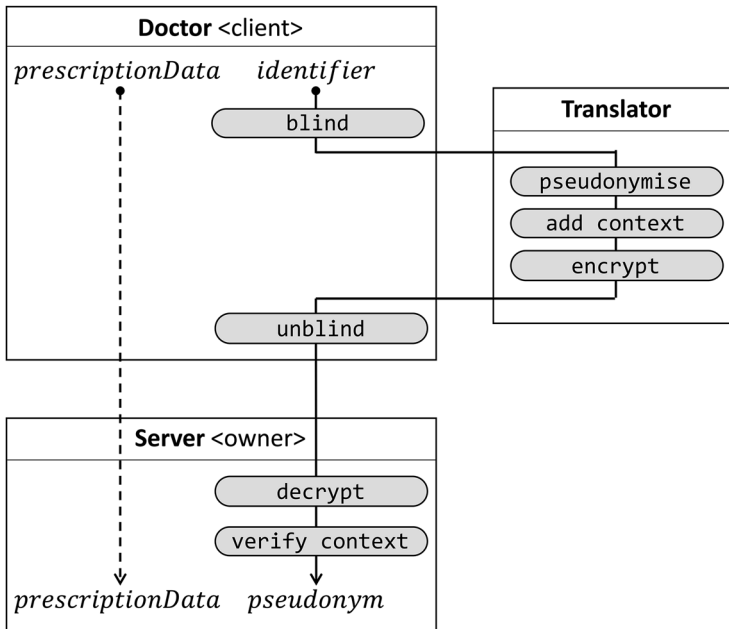


Fig. 5 Scenario 1 with high security

The use of ciphertexts also serves another purpose: protecting integrity and, hence, preventing malleability attacks whereby an attacker modifies unnoticed the ciphertext's context info.

The proposed scheme ensures the following properties:

- **Far-sighted client** learns citizen identifiers (=global identifiers) but no pseudonyms (=local identifiers).
- **Near-sighted owner** learns pseudonyms but no citizen identifiers.
- **Blind translator** learns neither citizen identifiers nor pseudonyms.

5.5 Pseudjoin

As previously discussed, in various *pseudjoin* projects, data sources have the capability to *autonomously* select relevant segments from their collection of personal data. The proposed translator could facilitate such projects. Figure 6 illustrates this concept for two data sources, although many more could be involved. Data sources S_1 and S_2 each possess distinct data, namely $data_{id}^1$ and $data_{id}^2$, related to the same citizen, known under identifier *id*.

Using the pseudonym, the collector can amalgamate data without ever directly accessing the citizen identifier. Therefore, data sources send information about

numerous citizens, along with encrypted pseudonyms, to the collector. The collector then decrypts the ciphertexts to obtain the pseudonyms and, based on them, consolidates records pertaining to the same individual, originating from various sources.

These *pseudojoin* projects typically operate as batch orders, not requiring real-time processing, unlike the prescription scenarios outlined previously. To optimize resource utilization, these projects could be scheduled with lower priority during off-peak hours when fewer real-time queries are processed.

When contrasted with existing practices described in Sect. 4.1, the proposed approach offers several advantages:

1. **Streamlined Communication Flows:** The approach features more elegant and simplified communication flows that closely align with business interactions. This streamlining benefits security and cost-efficiency.
2. **Reuse of an Existing Service:** Instead of building a dedicated service from scratch, this approach leverages an existing service. Reuse of components required by data sources or the collector (such as those for blind and unblind operations) results in additional security and economic benefits.
3. **Discrete Role of the Translator:** The translator is positioned on the sideline rather than at the centre.
4. **Enhanced Privacy:** Privacy is maximized by disclosing citizen identifiers only to the client and pseudonyms only to the owner (see Sect. 5.4). The translator remains oblivious to both and no longer learns personal data.

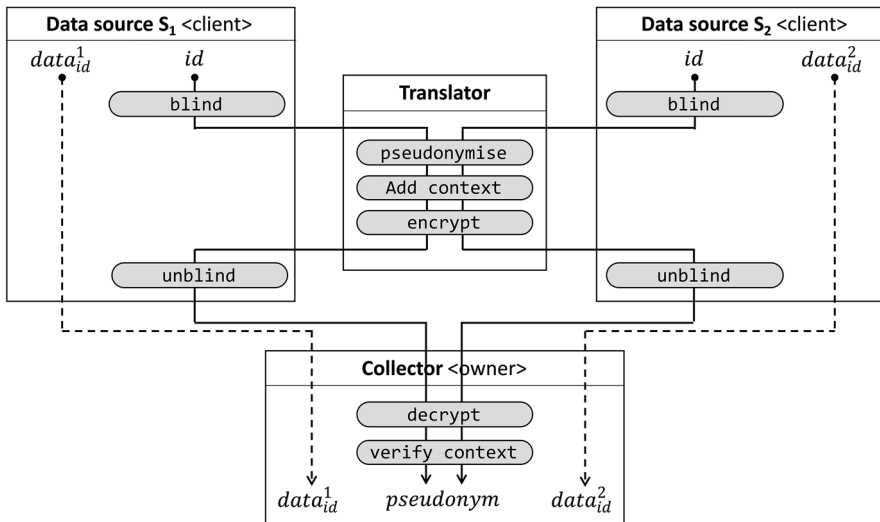


Fig. 6 Pseudojoining personal data with the help of the blind eHealth translator

6 LetheLink

LetheLink has been conceived by Smals in close collaboration with the KU Leuven. It is designed to address the issue outlined in Sect. 4.2, which focuses on *pseudo-joining* personal data scattered across various sources. In this context, at least one source does not have the capability to independently select pertinent records. The primary innovation here is the ability to achieve this while preserving a distributed and privacy-conscious setting. This means there is no reliance on a central translator and no risk of unintentional leakage of statistical or personal data to data sources or collectors.

In the scenario presented in Sect. 4.2, researchers should be able to access data represented by the unshaded areas in the Venn diagram (Fig. 2). Our current research is centred on revealing only the intersection of sets to the collector, with the prospect of expanding this scope in the future.

The fundamental concept is that every data source independently encrypts and pseudonymizes all *potentially* relevant information from its perspective before forwarding it to the collector. For example, in the case of Deliberation 20/020, SKR-FRC and IMA-AIM should provide encrypted and pseudonymized information about *all* cancer patients and *all* MS patients respectively. Due to prior cryptographic agreements between the data sources, the collector can only join and decrypt the information that belongs to the intersection (i.e., information about individuals who have both MS and cancer). As a result, *LetheLink* effectively prevents the leakage of sensitive personal or statistical data to the data sources or the collector, a feat that is challenging to achieve using alternative methods.

The collector's ability to decrypt information depends on the size of the intersection. A large intersection yields valuable insights, while a small result set may pose identification risks without significantly benefiting research purposes. In the latter case, the collector will be unable to join or decrypt *any* of the received ciphertexts, further reducing dependence on the correct behaviour of the collector.

The collector takes additional precautions to ensure data security and privacy. These include: (1) prompt deletion of any received but undecryptable ciphertexts; (2) execution of additional checks and operations on the decrypted data to detect anomalies and reduce re-identification risks.

LetheLink is revolutionizing the way *pseudojoin* projects are conducted. It introduces a consistent, highly automated process that eliminates the need for manual design, evaluation, and execution of data flows for every project. This automation accelerates project completion, leading to cost reductions. In essence, *LetheLink* not only enhances privacy but also optimizes the entire *pseudojoin* workflow, making it both efficient and cost-effective.

6.1 Distributed Pseudjoin Query

The primary objective is to join data originating from different data sources. We assume a *vertical partitioning* of citizen data; each data source possesses specific attributes related to individual citizens. For instance, one data source maintains attributes related to labour, another one healthcare-related attributes, yet another one handles pension-related information, etc. However, there is an essential exception to this partitioning: the shared use of a citizen identifier. Regardless of the specific attributes they manage, all data sources rely on this common identifier.

Imagine an academic researcher interested in non-aggregated —i.e., non-summarized— pseudonymized data. The data pertains to all workers who were born in 1990 or later, earn a salary above €50,000, and have self-employment as a secondary occupation. The researcher seeks further information about this specific demographic, particularly the types of diseases prevalent among these individuals, and specifics about their insurance coverage as self-employed persons.

This information can be retrieved by the query q shown below, where S_1 (national register), S_2 (department of taxes), S_3 (chamber of commerce) and S_4 (health care institution) are the involved data sources, managing attributes related to respectively citizenship, salaries, self-employment and healthcare:

q =Provide identifiers, agreed insurance data and agreed disease data of all citizens born not before 1990, with a yearly salary above €50,000 who are independent as 2nd occupation.

None of the data sources separately is able to answer this query. However, S_1 , S_2 , S_3 and S_4 can answer subqueries q_1 , q_2 , q_3 and q_4 respectively:

q_1 =Provide identifiers of citizens born not before 1990.

q_2 =Provide identifiers of citizens with a salary above €50,000.

q_3 =Provide identifiers and agreed insurance data of citizens who are independent as secondary occupation.

q_4 =Provide identifiers and agreed disease data of all citizens.

Each data source S_i , with $i \in \{1, 2, 3, 4\}$, executes subquery q_i locally. The result is denoted by $\mathbf{Recs}_i$, consists of records of the form $(id, data_{id}^i)$ and is provided as input to the *LetheLink* protocol. For instance, $data_{id}^3$ is the agreed insurance data for the citizen with identifier id . $\mathbf{Id}_i$ denotes the set of all the identifiers that occur in $\mathbf{Recs}_i$.

The protocol then pseudjoins its inputs in a distributed way. The output is denoted by $\mathbf{Recs}$ and consists of records of the form $(pseudonym, data_{id}^1, \dots, data_{id}^4)$. It is the pseudonymized result for query q . The protocol only retains records about citizens in the intersection of the Venn diagram in Fig. 6 and omits all the others.

The number of participating data sources is denoted by n . It equals four in our example; but can vary.

6.2 Privacy Properties

In the presence of a trusted entity, the protocol's functionality is straightforward: data sources share their data with it, which then computes the intersection of the datasets, replaces identifiers with pseudonyms, and forwards the resulting dataset to the collector. The *LetheLink* protocol is a distributed realization thereof; it eliminates the necessity for an absolute trusted entity. The privacy properties are defined as follows:

1. **No leakage to data sources.** Data sources do not gain access to new statistical or personal data through their participation in the protocol.
2. **No leakage to the collector.** The collector only acquires pseudonymised results (**Recs**) from the protocol execution, without gaining access to any other personal or statistical information.

In the proposed protocol, each data source (S_i) sends an altered version of their records (**Recs_i**) to the collector, wherein individual records are encrypted and pseudonymised. This approach leads to two distinct subproperties:

- **Conditional linkability.** The collector can link encrypted records of the same citizen provided by different data sources if and only if each set **Ids_i**, with $i \in \{1, \dots, n\}$, contains the citizen's identifier.
 - **Conditional decryptability.** The collector cannot decrypt a record in **Recs**, unless it is linkable.
3. **Minimum result size.** Although each subquery (q_i) may generate a large number of records, the resulting dataset from the main query (q) may be significantly smaller and contain only a few identifiers (i.e., the intersection indicated in grey in Fig. 7). This increases the identification risk while rendering data mining or statistical analyses meaningless. Therefore, a minimum size of **Recs** is agreed upon before the collector can access the results; decryption is only possible when that size is reached. (Due to word limits for this chapter, this property is not further elaborated.)

Without the *conditional linkability* property, detailed statistical information may leak, affecting the citizens' privacy. The collector would learn the cardinality—i.e. the number of elements—of each of the $2^n - 1$ regions of the Venn diagram. These detailed statistics about the population can be sensitive as such, but it also discloses detailed probabilities about individual citizens.

For example, if the collector is aware that a particular citizen belongs to the two left sets of the Venn diagram, it knows that this citizen belongs to either region **a**, **b**, **c** or **d**, with probabilities $|x| / (|a| + |b| + |c| + |d|)$ for $x \in \{a, b, c, d\}$, with $| \cdot |$ denoting the region's cardinality. This is more detailed than global statistics and is further aggravated when some of these regions are small or empty. In summary, the more the collector knows about a citizen, the more accurately it can guess other personal information.

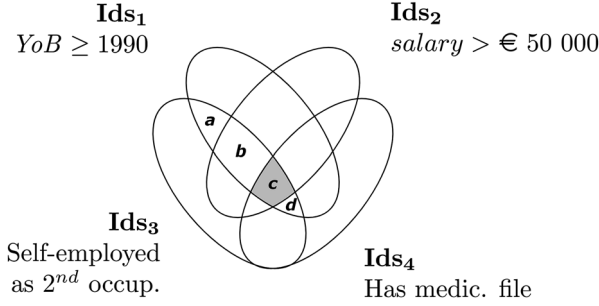


Fig. 7 Venn diagram with citizen identifiers for the example presented in Sect. 6.1

The complexity of this issue escalates as more data sources are involved. This is attributed to the exponential growth of the number of regions in a Venn diagram with increasing numbers of data sources. For instance, we have 15 regions for 4 data sources, 31 for 5 data sources, and 1023 for 10 data sources, and so on.

Due to the unique method of pseudonym creation, the cardinality of the $2^n - 1$ regions in the Venn diagram remains concealed from both the data sources and the collector. The collector only learns the size of the intersection in addition to the size of the individual n sets, which is typically non-sensitive and often publicly accessible information. In the rare instances where this poses a problem, a data source may add any number of fictitious records to solve this.

6.3 The LetheLink Protocol

This section presents the *LetheLink* protocol, abstracting from technical details and efficiency improvements. Essentially, each data source S_i has a set of cleartext records in the form of $(id, data_{id}^i)$ that it intends to deliver to the collector. Only if every data source provides an *encrypted* and *pseudonymised* record about the same citizen, the collector can *join* these records and *decrypt* them. Although intertwined in the *LetheLink* protocol, the two subprotocols are discussed separately: (1) *data encryption and conditional decryption* and (2) *pseudonymisation and conditional linking*.

6.3.1 Data Encryption and Conditional Decryption

For each relevant record $(id, data_{id}^i)$, data source S_i encrypts $data_{id}^i$ with a freshly generated secret key key_{id}^i (see Sect. 3.1), and sends the resulting ciphertext to the collector. As illustrated in Fig. 8, this key is split in n shares and each data source receives one share. Unless the collector (or another entity) can collect all the shares, it cannot recompose the original key. This only happens when it receives an

encrypted and pseudonymised record about that citizen from each of the n data sources. Indeed, along with this record, data sources also send n shares to the collector; one per data source.

In Fig. 8, data source S_1 splits key key_{id}^1 into three shares: $share_{id}^{1 \rightarrow 1}$, $share_{id}^{1 \rightarrow 2}$ and $share_{id}^{1 \rightarrow 3}$. It keeps the first share and sends the second and third one to S_2 and S_3 respectively. Similarly, S_2 and S_3 split a key and distribute the resulting shares over the data sources. If and only if a data source has a record for this specific citizen it sends to the collector a share-bundle, which is

$$shares_{id}^1 = (share_{id}^{1 \rightarrow 1}, share_{id}^{2 \rightarrow 1}, share_{id}^{3 \rightarrow 1}) \text{ for } S_1,$$

$$shares_{id}^2 = (share_{id}^{1 \rightarrow 2}, share_{id}^{2 \rightarrow 2}, share_{id}^{3 \rightarrow 2}) \text{ for } S_2, \text{ and}$$

$$shares_{id}^3 = (share_{id}^{1 \rightarrow 3}, share_{id}^{2 \rightarrow 3}, share_{id}^{3 \rightarrow 3}) \text{ for } S_3.$$

The first, second and third share in each share-bundle originate from S_1 , S_2 and S_3 respectively. With the first share from each of the three share-bundles, the collector recomposes key_{id}^1 , with which it decrypts the ciphertext received from data source S_1 . Similarly, with the second and third share from each of the three tuples, it recomposes keys key_{id}^2 and key_{id}^3 with which the collector decrypts the ciphertexts received from data sources S_2 and S_3 .

If at least one of the data sources does not provide a record about that citizen, the collector will be unable to decrypt any ciphertext related to that citizen, since it lacks at least one of the shares to reconstruct any secret key. Hence, we have realized the *conditional decryptability* property (see Sect. 6.2).

6.3.2 Pseudonymisation and Conditional Linking

For each citizen id , data sources build the corresponding pseudonym collectively by exchanging s -values, which are randomly generated large numbers. The citizen's data is delivered to the collector under this pseudonym.

A data source sends the same s -values to the other data sources if it has a relevant record about that citizen; in Fig. 9, S_3 delivers s_{id}^3 to S_1 and S_2 . However, if a data source does not have a relevant record about that citizen, it sends different s -values to the other data sources; in Fig. 10, S_3 sends $s_{id}^{3 \rightarrow 1}$ to S_1 and $s_{id}^{3 \rightarrow 2}$ to S_2 .

Data sources assemble a pseudonym for citizen id , based on the $n - 1$ received s -values and the locally generated one. This happens deterministically whereby individual s -values do not reveal any information about pseudonyms.

If all data sources deliver a record about citizen id to the collector (see Fig. 9), they all locally assemble the same pseudonym:

$$pseudonym = assemble(s_{id}^1, s_{id}^2, s_{id}^3).$$

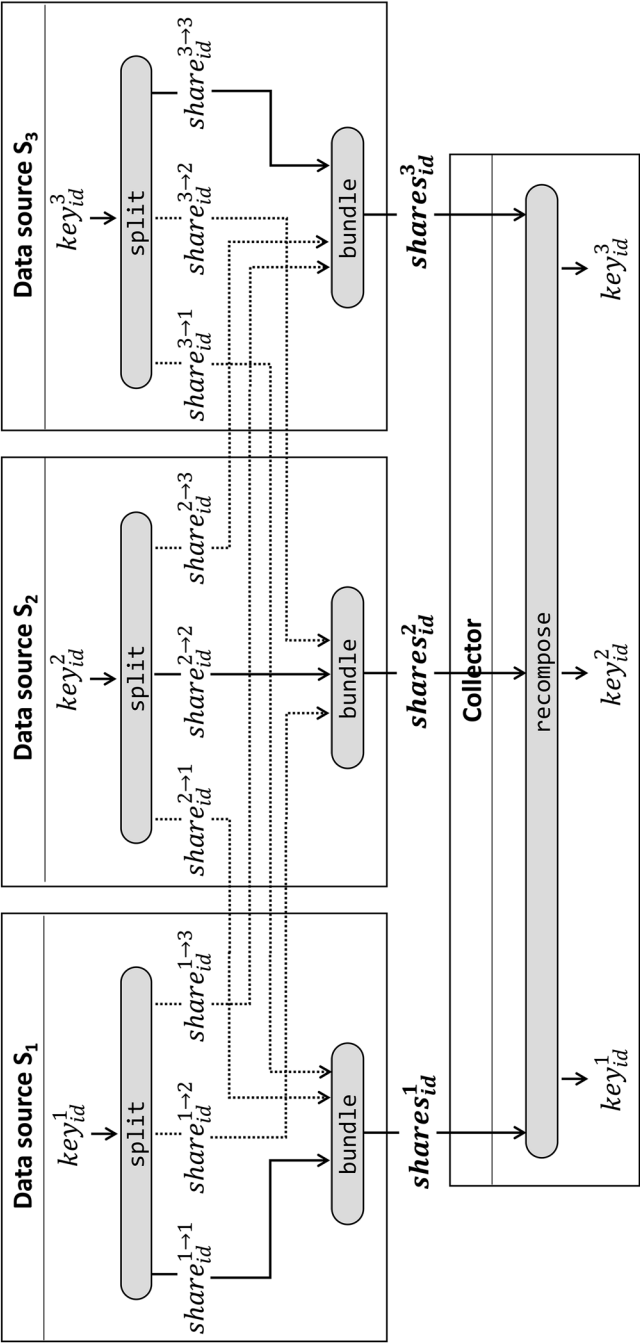


Fig. 8 Splitting secret keys into shares, distributing the shares and recombining the keys

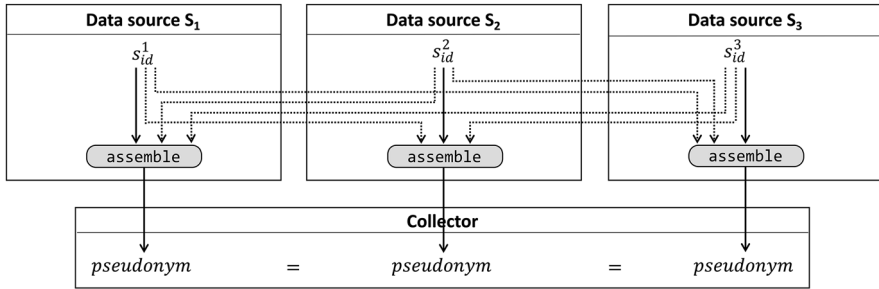


Fig. 9 Pseudonym assembly and delivery. ALL data sources have a relevant record about the citizen

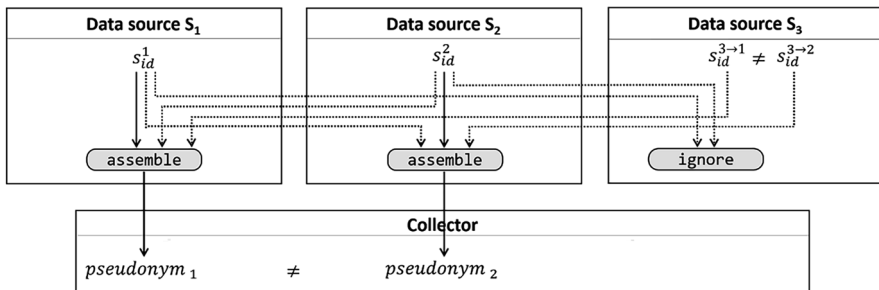


Fig. 10 Pseudonym assembly and delivery. NOT ALL data sources have a relevant record about the citizen

Every data source sends data under this pseudonym to the collector, enabling the latter to join records about the same citizen, received from different data sources.

If not all data sources deliver a record about citizen id , the collector ends up with different pseudonyms; in Fig. 10, data sources S_1 and S_2 deliver indeed data to the collector under $pseudonym_1$ and $pseudonym_2$ respectively:

$$pseudonym_1 = assemble(s_{id}^1, s_{id}^2, s_{id}^{3 \rightarrow 1}) \text{ and}$$

$$pseudonym_2 = assemble(s_{id}^1, s_{id}^2, s_{id}^{3 \rightarrow 2}) \text{ whereby } s_{id}^{3 \rightarrow 1} \neq s_{id}^{3 \rightarrow 2}.$$

This realizes the *conditional linkability* property, but there is still *leakage to data sources* (see Sect. 6.2). Colluding data sources learn for every citizen whether their s -values received from other data sources are equal or different. This reveals which citizens the other data sources provide data for as well as the cardinalities of the different regions in the Venn diagram.

Therefore, every data source sends a project-specific key to the collector, with which it probabilistically encrypts (see Sect. 3.1) each s -value before sending it to a data source. Hence, the assemble operations' input in Figs. 9 and 10 consists of one

unencrypted (full arrow) and two encrypted s -values (dashed arrows). The result is a (specially crafted) pseudonym ciphertext, decryptable only by the collector.

6.3.3 Synthesis

Per project, every data source S_i generates a secret project key $key_{project}^i$ and sends it to the collector. For each record $(id, data_{id}^i) \in \mathbf{Recs}_i$, S_i sends to the collector the tuple

$$(encNym_{id}^i, encData_{id}^i, shares_{id}^i)$$

with $encNym_{id}^i$ being the encryption of *pseudonym* (see Sect. 6.3.2) and $encData_{id}^i$ being the encryption of $data_{id}^i$ with secret key key_{id}^i . After decryption of $encNym_{id}^i$, the collector learns the pseudonym. If it has received a record with the same pseudonym from all other data sources, it derives the secret keys from the received shares (see Sect. 6.3.1), and decrypts $encData_{id}^j$ for $j \in \{1, \dots, n\}$, i.e., the encrypted data for all records with that pseudonym.

6.4 In Practice

The algorithmic and implementation complexity are hidden from the end-users. However, several actions and interactions at the business level are required for correct operation.

Before running the *LetheLink* protocol, the researcher contacts a coordinating entity. In Belgium, this could be the Belgian *Health Data Agency (HDA)* or the Belgian *Crossroads Bank for Social Security (CBSS)*, depending on the dominant type of the data. This entity knows what data each potential data source can deliver. The researcher engages in discussions with this entity to determine the minimal required data essential for conducting the research. This entity will typically assume the role of the *approver*, which compiles, signs, and publishes a description of the *pseudodojoin* project, but does not engage in the *pseudodojoin* protocol itself.

The business interactions, illustrated in Fig. 11, are described below:

1. Data sources and collector download the latest version of the *LetheLink* application (depicted by the screen with gears).
2. The approver creates and publishes a JSON³² file that contains all the information on how to execute the protocol, including security parameters, the n subqueries, a link to the deliberation (see Sect. 4), and the minimum result size.

³²JSON is an open, human- and machine-readable, standard file and data format.

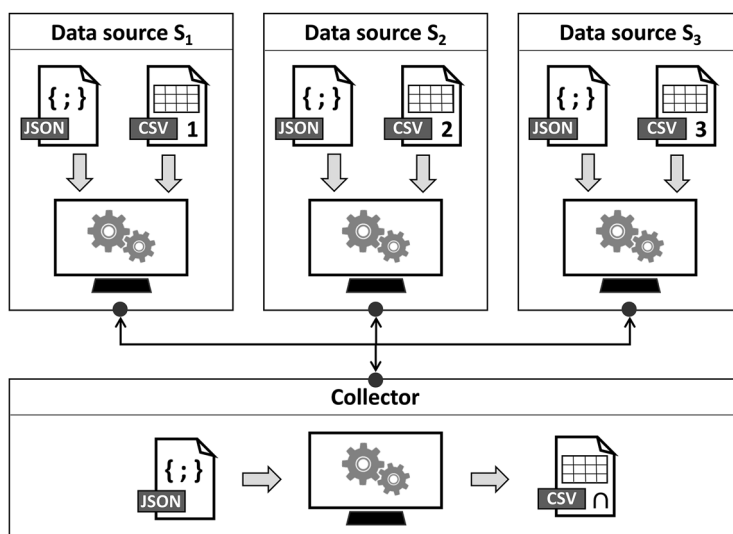


Fig. 11 Participants in *LetheLink* protocol execution, each with their local *LetheLink* application and its in- and output

3. Based on the deliberation or the JSON file, each data source independently generates a CSV³³ file containing the potentially relevant records with the requested *identified* personal data.
4. Data sources and collector initiate the *LetheLink* application, providing it with the JSON file. Data sources also input the CSV file.
5. The protocol is executed fully automated. If a sufficiently large number of citizens are involved in the result, the collector's application outputs a CSV file containing the decrypted, pseudonymised result.

Using *LetheLink* is convenient as there is no need to integrate with existing systems; it can be used stand-alone. This makes the approach highly suitable for on-demand *pseudoin* projects. However, prior legal approval to ensure regulatory compliance remains required.

7 Regulatory Compliance

The General Data Protection Regulation (GDPR) is the primary European regulation for protecting sensitive personal data. Key principles are *data protection by design*, *pseudonymisation*, and *data minimisation*.

³³ Comma-separated values (CSV) is a text file format that uses commas to separate values. A CSV file stores tabular data (numbers and text) in plain text, where each line of the file typically represents one data record.

7.1 Data Protection by Design

The solutions presented exemplify data protection by design, a concept defined by the European Commission as³⁴:

Implementing technical and organizational measures, at the earliest stages of the design of the processing operations, in such a way that safeguards privacy and data protection principles right from the start.

Data protection by design and default is embedded in Article 25 GDPR. Recital 78 of the GDPR further encourages this:

The controller should adopt internal policies and implement measures which meet in particular the principles of data protection by design and data protection by default.

7.2 Pseudonymisation

Article 4, (5) GDPR defines pseudonymisation as:

The processing of personal data in such a manner that the personal data can no longer be attributed to a specific data subject without the use of additional information, provided that such additional information (1) is kept separately and (2) is subject to technical and organizational measures to ensure that the personal data are not attributed to an identified or identifiable natural person.

Hence, replacing direct citizen identifiers, such as names and social security numbers, by pseudonyms does not necessarily result in pseudonymised data in the legal sense. However, it is a crucial element in a broader set of measures.

Both the translator and collector are data processors. Article 31, par. 1 GDPR states:

[...] The controller and the processor shall implement appropriate technical and organizational measures to ensure a level of security appropriate to the risk, including inter alia as appropriate: (a) the pseudonymisation and encryption of personal data; [...].

In the context of our eHealth translator, this means:

- The translator should protect the pseudonymisation key —i.e., additional data— and restrict pseudonymise, identify and convert requests to authorized entities.
- The collector should protect the combined dataset to mitigate re-identification risks. This may imply that researchers do not receive a full copy of the dataset.
- The data sources do not need to implement additional measures; they only learn encrypted pseudonyms with a limited longevity and do not manage additional cryptographic keys.

³⁴ See European Commission (n.d.).

In *LetheLink*, there is no translator. The additional data to re-identify is distributed over the data sources and the collector, which is a strong privacy-enhancing measure. However, measures are needed to avoid collusion between a data source and the collector; together, they can do substantial re-identification.

7.3 Data Minimisation

Pseudonymisation and *LetheLink* are examples of *data minimisation*, as they disclose only the minimal required data to the collector.

8 Conclusions

This chapter presented two innovations to *pseudjoin* (i.e., join and pseudonymise) personal data dispersed over multiple data sources, enabling researchers to gain novel insights that positively impact healthcare, policymaking, competitiveness, and society at large.

The *blind pseudonymisation service* is live and easily applicable for *pseudjoin*, projects. The existence of a translator makes it more suited for structural (controlled, long-term) exchanges. *Lethelink* is still experimental; its distributed character reduces infrastructural complexity but increases communication and software complexity. The absence of a translator makes it more suitable for on-demand (short-term) exchanges. It is especially suitable for projects where data sources cannot autonomously determine the pertinent records. In contrast to the other solution, it requires a vertical partitioning of personal data over the data sources.

Both solutions are promising for real-world pseudjoin scenarios. They exemplify privacy-by-design, successfully reconciling high privacy requirements with practical usability, while minimizing reliance on trusted third parties. This demonstrates how privacy-enhancing technologies help to comply with privacy legislation in the public sector.

Not only individual pseudjoin projects require a legal basis. Also, both presented solutions need one in order to permit their use for future pseudjoin projects. Approval of this legal basis requires sufficient trust in this solution.

In modern cryptography, algorithms and code are typically made public; the security only relies on the secrecy of cryptographic keys.³⁵ This transparency enables society to scrutinize and build trust in the proposed solutions. Nevertheless, due to their complexity, novel cryptographic solutions may face suspicion by non-expert stakeholders, who may see it as a counterintuitive black box. Bridging this gap between *theoretical* and *practical* transparency is challenging. Helpful were, among others, tailored clarification to key stakeholders and validations ('stamps') by esteemed academic partners.

³⁵ See Kerckhoffs (1883).

References

- Archer DW, Bogdanov D, Lindell Y, Kamm L, Nielsen K, Pagter JJ, Smart NP, Wright RN (2018) From keys to databases: real-world applications of secure multi-party computation. *Comput J* 61(12):1749–1771. <https://doi.org/10.1093/comjnl/bxy090>
- Bellamy FD, Fernandez AN (2023, November 15) A new era of privacy laws takes shape in the United States. Reuters. <https://www.reuters.com/legal/legalindustry/new-era-privacy-laws-takes-shape-united-states-2023-11-15/>
- Cavoukian A (2009) Privacy by design: the 7 foundational principles. Information and privacy commissioner of Ontario. Retrieved January 5, 2024 from <https://www.ipc.on.ca/wp-content/uploads/resources/7foundationalprinciples.pdf>
- de Montjoye YA, Hidalgo CA, Verleysen M, Blondel VD (2013) Unique in the crowd: the privacy bounds of human mobility. *Sci Rep* 3(1):1–5. <https://doi.org/10.1038/srep01376>
- de Montjoye YA, Radaelli L, Singh VK, Pentland AS (2015) Unique in the shopping mall: on the reidentifiability of credit card metadata. *Science* 347(6221):536–539. <https://doi.org/10.1126/science.1256297>
- eHealth Beraadslaging. (2013) Beraadslaging nr. 13/093 van 22 oktober 2013. eHealth. <https://www.ehealth.fgov.be/ehealthplatform/nl/sectoraal-comite/documenten?q=13%2F093>
- eHealth Beraadslaging (2019) Beraadslaging nr. 19/222 van 3 december 2019. eHealth. <https://www.ehealth.fgov.be/ehealthplatform/nl/sectoraal-comite/documenten?q=19%2F222>
- eHealth Beraadslaging (2020). Beraadslaging nr. 20/020 van 14 januari 2020. eHealth. <https://www.ehealth.fgov.be/ehealthplatform/nl/sectoraal-comite/documenten?q=20%2F020>
- eHealth Beraadslagingen. (n.d.) Beraadslagingen. eHealth. January 5, 2024 from <https://www.ehealth.fgov.be/ehealthplatform/nl/sectoraal-comite/documenten>
- eHealth Portaal (n.d.) Portaal van de diensten eGezondheid. eHealth. <https://www.ehealth.fgov.be/eHealth/Pseudonimiseren>
- eHealth Pseudonimiseren (n.d.) Pseudonimiseren & Anonimiseren. eHealth. <https://www.ehealth.fgov.be/ehealthplatform/nl/service-pseudonimiseren-anonimiseren>
- eHealth TherLink (n.d.) eHealth therapeutic links. eHealth. <https://www.ehealth.fgov.be/ehealthplatform/nl/service-ehealth-therapeutic-links>
- European Commission (n.d.) What does data protection ‘by design’ and ‘by default’ mean?. European Commission. https://commission.europa.eu/law/law-topic/data-protection/reform/rules-business-and-organizations/obligations/what-does-data-protection-design-and-default-mean_en
- Jensen M, Lauradoux C, Limnietis K (November 2019). Pseudonymisation techniques and best practices. European Union Agency for Cybersecurity (ENISA). <https://www.enisa.europa.eu/publications/pseudonymisation-techniques-and-best-practices>
- Kaaniche N, Laurent M, Belguith S (2020) Privacy enhancing technologies for solving the privacy-personalization paradox: taxonomy and survey. *J Netw Comput Appl* 171(2):1–32. <https://doi.org/10.1016/j.jnca.2020.102807>
- Kerckhoffs A (1883) La cryptographie militaire. *Journal des sciences militaires* 9:5–83
- Kruispuntbank Sociale Zekerheid (n.d.) Beraadslagingen. Kruispuntbank Sociale Zekerheid. <https://www.ksz-bcss.fgov.be/nl/deliberations>
- Lehmann A (2019) ScrambleDB: oblivious (chameleon) Pseudonymization-as-a-service. *Proc Priv Enhanc Technol* 3:289–309. <https://doi.org/10.2478/popets-2019-0048>
- Li Q, Wen Z, Wu Z, Hu S, Wang N, Li Y, Liu X, He B (2021) A survey on federated learning systems: vision, hype and reality for data privacy and protection. *IEEE Trans Knowl Data Eng* 35(4):3347–3366. <https://doi.org/10.1109/TKDE.2021.3124599>
- Marcolla C, Sucasas V, Manzano M, Bassoli R, Fitzek FH, Aaraj N (2022) Survey on fully homomorphic encryption, theory, and applications. *Proc IEEE* 110(10):1572–1609
- Morales D, Agudo I, Lopez J (2023) Private set intersection: a systematic literature review. *Comput Sci Rev* 49(1):1–24. <https://doi.org/10.1016/j.cosrev.2023.100567>
- Neubauer T, Heurix J (2011) A methodology for the pseudonymization of medical data. *Int J Med Inform* 80(3):190–204. <https://doi.org/10.1016/j.ijmedinf.2010.10.016>

- OECD. (2023). Emerging privacy-enhancing technologies: Current regulatory and policy approaches. OECD Digital Economy Papers, No. 351. OECD Publishing, pp 1–51. <https://doi.org/10.1787/bf121be4-en>
- Rannenbergh K, Camenisch J, Sabouri A (2015) Attribute-based Credentials for Trust. Springer eBooks. <https://doi.org/10.1007/978-3-319-14439-9>
- Rubinfeld J (1989) The right of privacy. *Harv Law Rev* 102(4):737–807
- Sciensano (n.d.) De technische architectuur van healthdata.be. eHealth. <https://healthdata.sciensano.be/nl/de-technische-architectuur-van-healthdatabe>
- Sweeney L (2000) Simple demographics often identify people uniquely. *Health* 671:1–34
- Verheul E, Jacobs B, Meijer C, Hildebrandt M, de Ruiter J (2016) Polymorphic encryption and Pseudonymisation for personalised healthcare. *Cryptol ePrint Archiv* 411:1–37. <https://eprint.iacr.org/2016/411>
- Verslype, K. (2023, October 16). Introductie tot de nieuwe eHealth pseudonimiseringsdienst. Smals Research <https://www.smalsresearch.be/basisprincipes-voor-een-moderne-pseudonimiseringsdienst/>
- Wieringa J, Kannan PK, Ma X, Reutterer T, Risselada H, Skiera B (2021) Data analytics in a privacy-concerned world. *J Bus Res* 122:915–925. <https://doi.org/10.1016/j.jbusres.2019.05.005>
- Zhao C, Zhao S, Zhao M, Chen Z, Gao CZ, Li H, Tan YA (2019) Secure multi-party computation: theory, practice and applications. *Inf Sci* 476:357–372. <https://doi.org/10.1016/j.ins.2018.10.024>

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.



Our Stakes in Data: How Do We (Re)Gain Democratic Control Over Digital Practices?



Barbara Prainsack

Abstract Against the backdrop of debates about the regulation of digital data and technologies, this chapter examines what needs to be done to (re-)gain democratic control over digital practices and infrastructures. Many of the harms arising from current digital practices neither break any laws nor contain open conflicts, yet they still harm important interest and stakes that people have in data. After analysing why this is the case, the chapter outlines two main ways to (re-)establish democratic control over digital data and technologies: The first is the enhancement of existing rights and the introduction of new rights suited to the digital era. The second is the reinforcement of democratic control over digital data and infrastructures. These goals cannot be achieved by merely adjusting existing regulatory instruments. Instead, they require a way of conceiving and recognising people's stakes in data that no longer rests on the classifications of the paper age. Such a new approach to data governance that enables effective democratic control would combine instruments of individual and collective control over data. The chapter concludes by arguing for the treatment of digital infrastructures as *public* infrastructures that require sustained public investment and effective democratic control.

1 The Problem: It's a Feature, Not a Bug

In 2021, Frances Haugen, a former product manager at *Facebook*, blew the whistle on the company's failure to make the platform safer for its users. Although the company knew that users were harassed, manipulated, or scammed on their site, it did not take effective steps against these practices if these measures could hurt their business

B. Prainsack (✉)

Department of Political Science, University of Vienna, Wien, Austria

e-mail: barbara.prainsack@univie.ac.at

© The Author(s) 2025

J. Goossens et al. (eds.), *Public Governance and Emerging Technologies*,
https://doi.org/10.1007/978-3-031-84748-6_7

131

interests. For example, *Facebook* reportedly granted high profile users all kinds of exceptions from their own rules—because banning them would reduce user numbers.¹

Haugen did not single out *Facebook* as the sole culprit. What she had seen there, she emphasised, was common practice also at many other social media enterprises and digital platforms. Neglecting the needs of users was a feature, not a bug: As long as digital platforms were owned and run by companies seeking to maximise their profits, with little or no public accountability and no systematic way of capturing and responding to different types of harm, these things would keep happening. In the words of Carly Kind of the Ada Lovelace Institute, ‘anything that doesn’t address the underlying business model of the data economy, which is acquiring data for the purposes of selling advertising space and maximizing engagement, would not fundamentally change the system’.²

We are changing the scene to Gurugram, India. Delivery worker Kumar (only his first name was reported) works for *Zomato*, a multinational food delivery enterprise headquartered in India.³ Kumar explains how he makes ends meet: *Zomato* grants him a base salary that is, however, not enough to live. To earn the extra money for him and his family to survive, Kumar tries to meet the company’s so-called incentive goals. To do so, he needs to be logged on to the *Zomato* app between 7 pm and midnight, and to complete every delivery, no cancellations allowed. The company also ‘motivates’ Kumar and his colleagues by publishing rankings of best performing delivery workers—a perk which does not come with a financial bonus but with the ‘honour’ of being recognised as a good employee.⁴

Both examples—although they are very different—are paradigmatic for the more subtle ways in which contemporary digital practices harm people. Many of these harms occur without the companies in question breaking any laws. The failure of digital platforms to protect users may be infringing all kinds of ethical rules, but only few of their practices are outright illegal. What happens to Kumar, for example, even takes place with his ‘consent’. In societies in which those at the bottom of the socio-economic pyramid are in the focus of data-driven surveillance, people like Kumar ‘agree’ to digital practices that harm them because they see no other choice.^{5,6}

Also for this reason, the power asymmetries between people and the corporations using their data is becoming ever more extreme. Moreover, tech companies have become rule-setting entities in our societies. Some of the largest players in this landscape—including Google, Amazon, Facebook and Apple, but also Baidu, Meta, or Xiaomi—have ceased to be market participants but become de facto market regulators, due to their size and influence.⁷ They create facts on the ground that regulators accommodate. The latter do so by generating result are generous tax incentives,

¹ See Horwitz and Seetharaman (2020) and Haugen (2023).

² MIT Technology Review (2021), p. 15.

³ See Barik (2021).

⁴ See Barik (2021).

⁵ See Draper (2017)

⁶ See also Prainsack (2015).

⁷ See Pasquale (2017).

antitrust regulations that remain toothless against multinational tech corporations, and data protection rules with unenforced sanctions, or with fines that remain below the pain threshold of large corporations. Tech companies have become a new Leviathan in this process: an *iLeviathan*⁸ that people submit some of their natural freedoms to receive something back that they consider essential. This is typically not physical security and the protection of their property, as was the case with the original, Hobbesian Leviathan, but the possibility to communicate across time and space, to buy goods in a faster and more convenient manner, to access public services that are only available online, and to use their time more effectively. For many, the latter is a necessary condition to be able to do all the other things they need to do to hold a job and care for their families.

This situation reflects a massive deficit of democratic control over digital data and infrastructures.⁹ Democratic control here comprises all practices and measures that help to ensure that people can influence decisions that affect them, including economic ones.¹⁰ It includes established forms of democratic representation and control, but also contributions to policy making and democratic control coming from civil society institutions such as public media, human rights advocates, or citizen associations. It is the form of democratic action that has not been captured by tech industry interests. An extreme manifestation of our collective lack of democratic control over digital data and infrastructures is that public actors increasingly rely on technologies owned by for-profit corporations for ever wider aspects of public service provision.

If the loss of democratic control over digital technologies is such a big problem, why has it not been solved? If it harms the interests of so many people, why have countries—at least those that are democracies—not issued laws and policies to change the status quo? The next section will discuss some of the reasons for this.

2 Why Has the Problem Not Been Solved?

2.1 Lobbying

One obvious reason is lobbying. In recent years, technology corporations have invested large sums of money in making their interests heard. Between 2014 and 2017 alone, big tech interests lobbying in Europe increased by 278%.¹¹ The biggest spender at the time, *Google*, shelled out 4 m Euros per year on lobbying when the study was carried out. Tech corporations are also known to resorted to other, less tangible strategies of making their interests heard, such as offering former

⁸ See Prainsack (2019).

⁹ See EGE (2023).

¹⁰ See Goodin (2007) and see also Cumbers (2020).

¹¹ See Kergueno (2022) and see also Cooper and Hirst (2017).

politicians lucrative positions or board memberships, or hiring staff from EU institutions.¹² At the end of the 2010s, four out of seven lobbyists accredited with the European Parliament had been hired directly from the Parliament to lobby for their new employers.¹³ Research has confirmed that businesses with deep pockets have better access to the Commission than less well-endowed enterprises, citizen groups, or professional associations.¹⁴ Although access does not equate results,¹⁵ it is plausible to assume that this situation places (large and wealthy) businesses at an advantage vis-à-vis organisations championing civil rights and other public interests.¹⁶

2.2 *Imaginative Captivity*

Next to some people's vested interest in retaining the *status quo*, there is also a much less visible reason for our collective inability to create policies that realise an equitable distribution of harms and benefits in the digital era. In our book *The Pandemic Within*, Hendrik Wagenaar and I discuss a form of imaginative captivity that prevents us from looking beyond our immediate moral, and practical concerns.¹⁷ Because we have been socialised into specific systems of practices and judgments, the features of these systems have become self-evident to us. They comprise specific methods of appreciating, seeing, and acting. Their shape and meaning are built into the very structure of our language.

What makes it difficult to break free from this imaginative captivity is that we cannot merely think ourselves out of it. It is anchored in our personal, social, economic, and political practices. We are so used to digital platforms using data in unforeseeable ways, and us having no possibility to change it, that many of us have resigned to this. We accept it as a fact—as part of how the world is. What is more, we enact this reality anew every day and become complicit with it. It becomes 'our' reality, which makes it even more difficult to act and think upon it differently.

2.3 *The Frame of Fair Market Competition*

Another reason why it has been so difficult to put effective measures to ensure a more equitable digital society on the policy agenda is a specific way of framing the problem that is to be solved. In many world regions, including the European Union

¹² See Freund and Bendel (2017).

¹³ See Freund and Bendel (2017).

¹⁴ Dür and Mateo González (2012), pp. 983–984 and see also Ruohonen (2019).

¹⁵ See Dür (2008) and see also Dür (2019).

¹⁶ See also Atikcan and Chalmers (2019).

¹⁷ See Wagenaar and Prainsack (2021) and see also Prainsack (2023).

(EU), the dominant frame for the regulation of digital technologies is fair market competition. Although individual instruments such as the General Data Protection Regulation (GDPR) or the Artificial Intelligence (AI) Act go beyond this, the focus of regulatory efforts has mostly been on the services offered by digital platforms, and not the social and material infrastructures that enable the generation use of digital data.

In recent years, the harmful concentration of power with a few powerful platforms has been recognised as a problem by policy makers around the world. The dominant response, however, has been to adopt new (*ex ante*) regulations and revisions to existing legal frameworks, especially in the realm of competition law. Within the EU, the Digital Markets Act (DMA), in conjunction with the Digital Services Act (DSA), plays a central role in the Union's strategy for a unified digital market. It is meant to respond to the problem that regulatory efforts at the Member State level may fall short in addressing competition and unfair practices, which could potentially fragment the Internal Market.¹⁸ Also here, it is apparent that the focus is on regulating services, and not on the underpinning infrastructures and other structural aspects of the economy.¹⁹

3 What Should Be Done?

What can be done to change this situation? As colleagues and I argued elsewhere,²⁰ merely giving people more individual-level control over their data is not sufficient. The data individualism of the past decades has even been an obstacle to more effective (and more collective forms of) control and oversight over datafication and data use. In the following section, I will explore what is needed to (re)gain democratic control over digital practices. Such democratic control can take place at various levels: From communal to regional to national, encompassing not only elections but also public policy making in the widest sense, including forms of public accountability and oversight. Exercising democratic control includes the protection of people's rights and interests. I will first describe what new rights may be needed, and then explore how 'old' rights require new policies instruments to protect them. I will conclude by arguing that digital infrastructures should be recognised as public infrastructures and funded, owned, and governed as such.

¹⁸ See Akman (2021).

¹⁹ The explicitly stated objective of the DMA is 'to ensure a contestable and fair digital sector in general and core platform services in particular, with a view to promoting innovation, high quality of digital products and services, fair and competitive prices, as well as a high quality and choice for end users in the digital sector' (DMA n2, Recital 79; cf. Akman (2021), p. 6). Akman concludes that the 'two concerns that drive the DMA are, therefore, contestability and fairness on digital markets' (Akman (2021), p. 6).

²⁰ See Prainsack (2022a, b).

3.1 New Rights: A Right to Human Contact, and a Right Not to Be Legible

Strengthening democratic control over digital data and infrastructures needs to start with the rights and needs of people. As the European Group of Ethics noted in their Opinion on Democracy in the Digital Age,²¹ the needs and rights of people need to be protected in ways that are adequate to the digital age, and not with instruments from the paper age. Sometimes this requires making sure that specific rights—such as the right to privacy—are realised with new instruments. Sometimes it also requires introducing new rights altogether, such as the right to human contact.

Human contact is a basic need of people. It is threatened by several developments, from automated medicine to artificial intelligence in public and commercial service provision to the digitalisation of education. It is thus not far-fetched to assume that without policies making sure that this does not happen, the masses—and especially, poor people—will interact less with human service providers and more with automated services. In such a scenario, children who are ‘disruptive’ at school may no longer receive support from humans but go to school in the meta-universe.²² Those who cannot afford a subscription with a boutique medicine clinic will have access to a human medical professional whenever they need it, while the rest may have to do with automated services—except in cases of acute or serious diseases. And those without disposable incomes or other sources of wealth will pay with their data for services that the rich pay with money.²³

To address these problems, in 2017, a Report by the Dutch Rathenau Institute²⁴ suggested to introduce a new human right to meaningful human contact. The Rathenau report positions this right in the context of ‘the robot age’. This right should arguably be more broadly seen as a fundamental right in the age of automation. It should be understood as an entitlement of individuals to interact with humans if they so desire and whenever important interests and needs are concerned. While it would not extend to the right to order a pizza with a human, it would ensure that when dealing with financial matters, healthcare, or public administration, a human would need to be available within a reasonable timeframe if requested by a person.

Another human right that should be added to the list of rights in the age of digitisation and automation is the right not to be legible outside a narrowly defined range of contexts where such legibility is necessary for public administration and the provision of public services. This proposed right takes inspiration from James C. Scott’s notion of legibility in the context of public administration—but expands this analysis to all situations where people are made legible by organisations, may they be governments, corporations, or other entities. Scott (1998) argued that states and other powerful institutions often seek to simplify and standardise the

²¹ See EGE (2023).

²² See Haugen (2023).

²³ See Prainsack and Forgó (2022).

²⁴ van Est et al. (2017), p. 44.

information they have about individuals and communities to better control and govern them. In the context of statecraft and governance, legibility means the state's ability to collect information about its population, resources, and territory in a way that is easily understandable and manageable—via censuses, land surveys, standardised identification systems, and other data-gathering methods. While legibility, Scott notes, may appear beneficial in terms of enabling efficient administration and planning, it also leads to the oversimplification of complex social realities. By reducing people and communities to quantifiable data points, important local knowledge and practices that do not fit into standardized categories are often overlooked or dismissed.

The consequences of such *legibility by default* can be detrimental to both the state and the people it governs. When the state imposes rigid and uniform templates onto diverse and intricate social structures, it fails to account for the unique needs and complexities of different communities. This can result in policies that are misaligned with local realities, leading to social unrest, resistance, and even policy failures. In this context, Scott highlights the importance of *metis*, which refers to local, practical knowledge and skills acquired through experience and passed down through generations. This form of knowledge often differs from formalised, abstract knowledge endorsed by states and experts. *Metis* could be an alternative to standardised data collection. For example, rather than the sex of a child being registered centrally at birth, knowledge of a person's gender identity—or sex, where necessary—can be obtained individually in different contexts on a need-to-know basis.

Also here, the Dutch Rathenau Institute has played a pioneering role. Already in 2017, it proposed a right not to be measured, analysed, or coached. It considers this right as a fundamental aspect that should be recognised and protected under international law. In particular, the Report recommends that the Council of Europe clarifies to what extent 'the right to respect for privacy implies the right to not be measured, analysed or coached'.²⁵ In addition to these new rights, we also need new institutions and ownership structures that will create the conditions of possibility for these rights to be exercised.

3.2 Democratic Control Over Digital Infrastructure and Digital Data: What Regulation Can Learn from Trains and Cars

In July 2023—less than a year after Musk's acquisition of Twitter, and amidst agitated debates about the dangers of generative AI, Carissa Veliz voiced her frustration with the 'pinkie promises' of AI companies vowing to better manage AI risks.²⁶ The world did not need any more voluntary commitments, Veliz argued. Instead, there

²⁵ Ibid.

²⁶ See Veliz (2023).

are important lessons to be learned for regulating AI from the history of regulating the car industry. One is that industry will typically lie about what is possible to for them to do in terms of the safety, and they will blame users for any harms resulting from their product. Another lesson to learn, Veliz argued, is the importance of well-funded enforcement for effective regulation. The billions that get lost for public coffers each year because of the understaffing of tax offices should serve as a warning in this respect. As if to prove Veliz' points right, in the same month, *Google* made headlines by calling upon the Australian government to relax their copyright laws to make it easier to train AI. Australian law, company representatives argued, should also become more flexible about 'sending sensitive data offshore'. Following Europe's example in regulating AI was a bad idea—and doing so would mean to risk losing talent and opportunity. 'Companies should not face too onerous requirements to explain how AI reaches its decisions', *Google* argued.²⁷

What happens if we push the car industry analogy further? For example, will greenwashing, and companies trying to cover up their mistakes, be part of the trajectory that tech companies will follow? If so, how can we avoid that history repeats itself in this respect? Perhaps the car industry is an instructive template indeed. When it comes to mobility, one shift has been to go beyond the individualistic paradigm and place more emphasis on sharing and collective solutions. Another one has been increased emphasis on accessibility and inclusivity, and the integration of mobility with sustainability goals. And then, there is the question of whether we should have cars at all. While some believe that cars are here to stay—albeit in more affordable, 'cleaner', and safer iterations, others believe that we should focus on other forms of transportation. Max Fisher (2022) took such a stance regarding social media: The latter have harmed societies to such an extent that they should be treated more like tobacco products, Fisher argued. Rather than trying to get industry to create better, safer products, regulation should focus on limiting access.

Could this argument be applied to digital technologies more broadly? Digital technologies cover a much wider range of practices than penetrate virtually all aspects of our societies. In recent decades, ever wider aspects of our lives and bodies have been datafied and digitised. Digital infrastructures, I argue, have become so central to the functioning of our societies that they should be recognised as public infrastructures.

3.3 *Digital Infrastructure as Collective (Common) Property*

In addition to learning from the regulation of the car industry, when it comes to how we should treat digital infrastructures, we can learn from the history of another key infrastructure, namely railways. The first railroad systems were built in the early nineteenth century. These initial railroads were often owned and operated by private companies or entrepreneurs. From the mid-nineteenth century to the early 20th,

²⁷ See Smith (2023).

railroads expanded rapidly as they became crucial for industrialization. Many nations saw a mix of privately owned railroads and government involvement in building and regulating them. Governments began to acquire or regulate railroads to ensure public access and prevent monopolies. In the mid-twentieth century, many governments nationalized railroads as part of broader efforts to manage transportation infrastructure, and to improve efficiency and coordination. In the second half of the twentieth century, to increase competition, efficiency, and innovation, many governments sold off their state-owned rail assets. More recently, some countries have adopted public-private partnerships, where governments collaborate with private companies to finance, build, operate, and maintain rail systems. This approach claims to combine the strengths of both sectors and to lead to improved infrastructure and services. The reality is a bit less sanguine, with the public part of these partnerships often funding private profits.

If we consider digital infrastructures as essential public infrastructures, in an analogy to railroads, the focus of our debate shifts. Like railways, digital infrastructures were made possible by a cluster of technological advancements that coincided with economic changes (some of which they contributed to). Their development can be described as a patchwork of public and private initiatives enabled by a range of existing public infrastructures. Both have played an essential role for a wide range of economic and political activities, from trade to communication to elections. Both form part of critical infrastructure.

If, despite all the differences, we accept that there are sufficient similarities between railways and digital infrastructures, then the lessons to be learned are obvious. The prioritization of profits over other considerations such as safety, service, quality, and long-term infrastructure investment, in the case of privatized railways, has led to the neglect of less profitable routes, reduced frequency and reliability, and higher fares—especially in cases where governmental regulation or the lack of competition allow rail companies to control pricing. In some cases, privatization has also led to *de facto* rail monopolies, limiting competition and potentially leading to higher fares and reduced service quality. In systems with multiple private operators, coordinating schedules, ticketing, and infrastructure improvements can be complex and challenging. If a private rail company faces financial difficulties, this can impact the stability of the entire rail network, leading to service disruptions and the need for government intervention to prevent collapse. Finally, the lack of democratic control makes it more difficult to secure long-term investments and to align the operation of the railroads with other public interest goals.

Digital infrastructures are different from railways regarding the proportion of tangible assets, their embeddedness in local and regional scales, and their environmental impact, for example. But they both are infrastructures that are critical for the functioning of societies and should thus be jointly owned and governed. The ownership part of this equation is not trivial. To ensure that digital infrastructures are used in such a way that they facilitate and support activities that benefit the wellbeing of people and the planet, and increase, rather than decrease, social justice, they would need to be owned and governed by people—ideally in the form of a state property regime, or even in a commons property regime (*res communis*) in analogy to local

commons (see next section, Layer 2). The advantage of a communal (‘commons’) property regime would be that the common-pool resource would be owned by all people—here, the residents of a country—together. While they would not necessarily be governed by means of direct democracy, the governance model would be decided by in a deliberative way, and all those who cannot, or do not want to, be involved directly in governing would give their ‘consent to be governed’,²⁸ rather than this consent to be presumed.

3.4 *Three Layers of Stakes in, and Control Over, Data*

In the remainder of this chapter I will argue that democratic control over digital data—including over decisions on what aspects of our lives and bodies are datafied, and what purpose these data are used for—can be conceived as taking place at three (overlapping) layers (see Table 1): The first layer realises individual-level control of people over whether and how ‘their’ data is collected and used. The second layer uses collective instruments of control and oversight, often at the level of communities and regions. The third comprises national-level control over digital data and infrastructures. Each layer captures different stakes that people have in the phenomena and knowledge that is captured in data, ranging from the most intimate (Layer 1) to very stakes (Layer 3). I will discuss each of them in turn.

3.4.1 **Layer 1: Individual-Level Control of People Over ‘Their’ Data (and a New Definition of *Personal Data*)**

It is largely uncontroversial to state that people should have meaningful control over their data. But what data is ‘their data’? The answer to this question is not trivial. In many legal systems around the world, personal data are those which refer to an identified or identifiable individual. They are typically seen to deserve a higher level of protection than data that do not relate to specific individuals. Within the category of personal data, in turn, certain types of data—such as data on health, sexual

Table 1 Layers of democratic control over digital technologies and practices (source: author)

Layer 3:	All other data	National-level control
Layer 2:	Important (non-intimate) stakes in phenomena and knowledge that are datafied	Community-level control, oversight, and ownership
Layer 1:	Intimate stakes in phenomena and knowledge that are datafied	Individual-level control

²⁸ See Koenig (2014).

orientation, race, or union membership—have been labelled particularly sensitive, and many data protection frameworks have special protections in place for this category of data. Separating out only certain types of data as needing more protection than others, however, is anachronistic. Today, any dataset—if linked with suitable kinds and quantities of other datasets—can lead to the identification of the person they came from. The distinction between personal and non-personal is thus becoming increasingly brittle.²⁹ Moreover, almost any type of data can also be used to make inferences about other people than those that the data come from. These inferences can be based on seemingly innocuous pieces of information about a person, such as whether they buy sports goods online, or whether they have a telephone landline. Even non-human data can cause harm to people. For example, if environmental datasets showed higher levels of carcinogenic toxins in a specific area, then people living in this area could face discrimination from insurers based on this information.³⁰ In this sense, a better definition of *personal* data than the one that most data governance frameworks use today would be to refer to data that pertain to all phenomena and entities that people have a strong stake in (i.e. that they have an important relation to). For most people, this would include data about their bodies, their homes and other environments of daily living, and data on their immediate family. For some people and in some communities, the notion of personal data would expand also to the animals, plants, and rocks that form our immediate environment, and to human ancestors. In a data governance framework corresponding with such a cosmology, environmental data can be *personal* data just as health records are.³¹ Rather than having a legal definition of personal data as those data that refer to an identified or identifiable human, such a new understanding of personal data would require that data governance takes into account the context of specific instances of data collection and data use and the stakes that different people and communities have in the information and knowledge that is datafied.

Moving from the question of how personal data should be defined to how control over such data should be exercised, at the most intimate level, where data concerns people's bodies and immediate personal lives (see Fig. 1), knowledge should be datafied and used only with people's explicit, voluntary, and meaningful consent³²—unless this is not feasible or not meaningfully possible. Such consent, in turn, can be considered voluntary and meaningful only if alternatives are available, meaning that those who do not agree to the datafication, or data use do not suffer undue disadvantage. For example, a person who would not want to have a heart monitor implanted that sends data directly to the device manufacturer or to the hospital³³ should be

²⁹ See Purtova (2022).

³⁰ See Prainsack et al. (2022b).

³¹ See also Kukutai and Taylor (2016); Chao (2018); Yunkaporta (2019); Cormack and Kukutai (2022).

³² Informed consent, here, is not to be understood in a narrow, technical sense, but in a sense as the process by which people are informed about how data are intended to be collected and used and agree to these uses (either as a one-off event or a process).

³³ See von Arx 2023.

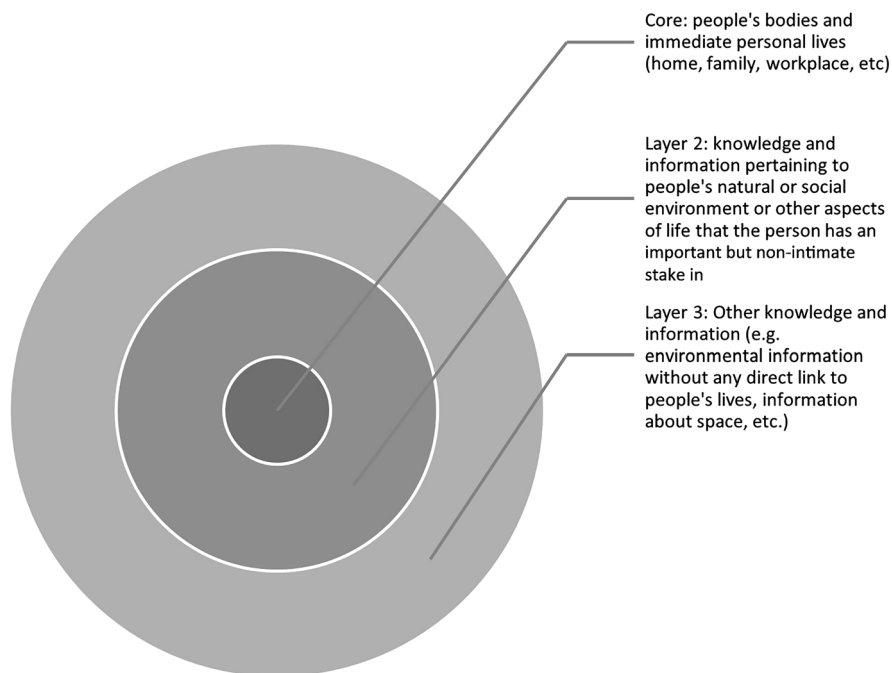


Fig. 1 Three layers of control over data and the phenomena that they are seen to represent (source: author)

offered a reasonable alternative if there is one available. Similarly, people should be able to use online services without being tracked if this is important to them.

In current data governance frameworks, as almost all data in this first layer would be classified as personal data, many countries already require the informed consent of the data subject for the use of personal data in most situations. In an ideal situation, however, the notion of the personal would become more flexible, paying attention to what information and knowledge people *actually* have a stake in, instead of whether the data can lead to their identification.

3.4.2 Layer 2: Community-Level Control, Oversight, and Ownership

The second layer (Fig. 1) comprises data representing information and knowledge that do not pertain to a person's immediate intimate personal life, but that they, nevertheless, have an important, non-intimate stake in. Examples include information about a person's natural and social environments, or their ancestors. Collective instruments of control, oversight, and ownership are best equipped to protect the stakes that people have in these data. Collective instruments, in turn, comprise all measures by which people enact control and oversight over data not as individuals but as part of a collective. They range from policy making at the level of legislation

(via representation in national parliaments or other legislative bodies, see below, layer 3) to exercising collective control and ownership in the form of a commons. Commons are best understood as social institutions for governing common-pool resources³⁴ that are guided by the values of fairness, equality, justice, and sustainability.³⁵ They are characterised by ‘a plurality of people (a community) sharing resources and governing them and their own relations and (re)production processes through horizontal doing in common, commoning’.³⁶ Rather than being anti-state, or replacing individual control, they represent a middle ground between state control and atomistic individualism. As an element of a complex system of polycentric governance,³⁷ commons enable self-governance of people via informal rules, social norms, and more formal procedures³⁸—as an alternative to both hierarchical-bureaucratic government by state organisations,³⁹ and to the idea that people should exercise control over the data that they have a stake in only at the individual level. As the example of Kumar at the beginning of this chapter showed, leaving such decisions with individuals ignores that individual ‘choice’ is often not free but shaped by important needs and structural constraints.

Based on longstanding empirical studies, the political scientist Elinor Ostrom developed ten ‘design principles’ for the governance of common-pool resources. To start with, to be open to the possibility of good governance, commons need to have clearly defined boundaries. This pertains both in terms of physical resources and regarding the owners of the commons. Applied to digital commons, it needs to be clear what data, infrastructures, and other material or immaterial resources are included, and who the human members of such commons are who co-own the resource and govern it together. Second, governance rules need to be appropriate to what Ostrom calls ‘local’ conditions, e.g., to the ecological conditions that the resource is part of. Commons around financial data, thus, would need to have governance rules that fit the field of finance. There also need to be mechanisms for inclusive collective decision making, and processes must be in place for monitoring appropriators. In case of rule violations, graduated sanctions need to be applied, and conflict resolutions must be fast and inexpensive. Importantly, the self-governance of the commons needs to be recognized by the political and legal system that the commons is surrounded by, so that nobody can interfere with what the members of the commons decide. Finally, governance should be organized in layers of nested enterprises.⁴⁰

³⁴ Common-pool resources are any naturally occurring or human-made materials (and immaterial assets) from which it is very difficult but not impossible to exclude users.

³⁵ See Bollier and Helfrich (2019).

³⁶ De Angelis (2017), p. 10 and see also Prainsack et al. (2022b), p. 10.

³⁷ See Ostrom (2010).

³⁸ Bollier and Helfrich (2019), p. 119 and see also Wagenaar and Bartels (2024).

³⁹ Hollo (2022), p. 122.

⁴⁰ See Ostrom (2010) and see also Agrawal (2007).

Although some forms of collective control over data—including data commons and cooperatives—already exist in our societies today, they are not yet used to their full potential. Moreover, the involvement of industry representatives and the corporate capture of policy makers has meant that policies in most countries, rather than protecting and enabling commons, have facilitated commercial profiteering from people's data. Often this has come at the cost of the people who have stake in the data or the things that they are seen to represent. For law to regulate datafication and data use effectively and with public value in mind would require a much stronger commitment to public welfare. This would, among other things, entail a reorganisation of the production and exchange of goods and services—e.g. by making stronger use of commons to do so. Commons can emerge bottom-up from and by the initiatives of citizens. To be able to flourish and to survive, however, they require laws and policies to protect them against commercial takeover, and to ensure that commoners have access to financing. That would be one of the novelties compared to the status quo: That legal frameworks are created to protect and support commons. People who want to set up commons need the financial means to do so, just as people who set up to start a commercial business. Moreover, law and policy need to ensure that commons can complement other, publicly and commercially provided goods and services, without replacing the former and being cannibalised by the latter.

3.4.3 Layer 3: National-Level Control Over Data and Infrastructures

The third layer contains all information and knowledge that is not included in layers 1 and 2. It is information and knowledge that people do not have an immediate stake in, such as information and knowledge about the wider environment, space, or the cultural heritage of a group that they do not identify with. Control over this kind of information and knowledge can be mediated through forms of collective governance at higher levels.

One such higher level of control is legislation and other public policy to ensure that data use that poses high risks is effectively outlawed, and data use that does not entail such high risks but that is likely to create high public value should be proactively supported—by easing regulatory burdens for such data use, or by providing public funding (i.e. in the form of hands-on support with getting permits, preparing data files, etc). The approach of solidarity-based data governance (in short: data solidarity) could be used to realise these principles.⁴¹ Data solidarity is based on three main pillars: Pillar I comprises instruments and measures that support and facilitate data use that is likely to create significant public value without exposing individuals or groups to significant and undue risks. Pillar II focuses on preventing and—where preventing has not succeeded—mitigating harm emanating from data use. Lastly, Pillar III aims to ensure that a fair proportion of the profits that emerge from commercial data use come back to the public domain that has enabled the data use via the data work of individuals and public infrastructures in the first place.

⁴¹ See Prainsack et al. (2022a, b).

4 Conclusion

There are lively debates over how to best regulate digital technologies and digital data, to ensure that they are used for the benefit of people, and to reduce the harm stemming from digital practices. So far, the main approach has been to increase the control that data subjects have over the use of their data at the individual level. This focus on individual agency—in combination with an attempt to ensure fair market competition—is, however, unsuited to address the structural issues that lie at the core of data harms, such as the concentrations of power with large tech corporations. As a result, the risks and benefits emerging from digital practices are distributed very unequally across and within countries.

Trying to answer the question why laws and policies have not, thus far, managed to effectively increase democratic control over datafication, data use, and digital infrastructures, I offered three explanations: Lobbying, imaginative captivity, and the dominance of the frame of fair market competition. Taken together, they lead to a cluster of ideas and practices that is very difficult to change—unless we conceive of people's stakes in digital practices, and data, in a different way. There are two main avenues to do this. The first is to ensure that our protection of existing rights and needs is adequate to the digital era and the creation of new rights, such as the right to human contact, and the right to be illegible. The second avenue is to strengthen democratic control over digital data and digital infrastructures, not merely by readjusting existing regulatory instruments but by changing the way we categorize data and our takes in them. Starting with infrastructure, I argued that digital infrastructures should be recognized as public infrastructures and governed as such; the history of governing railways provides an instructive example. Regarding the expansion of democratic control over digital data, I proposed an approach that combines different levels and instruments individual and collective control over data depending on the quality and intensity of people's stakes in the phenomena that are datafied.

Acknowledgements I am grateful to Antonia Stanojević, Jurgen Goossens, Stephan Grimmelikhuisen, Esther Keymolen, Marcel Robeer, and Hendrik Wagenaar for comments on earlier drafts of this manuscript. The usual disclaimer applies.

References

- Agrawal A (2007) Forests, governance, and sustainability: common property theory and its contributions. *Int J Commons* 1(1):111–136
- Akman P (2021) Regulating competition in digital platform markets: a critical assessment of the framework and approach of the EU digital markets act. *Eur Law Rev* 47(1):85–114
- Atikcan EÖ, Chalmers AW (2019) Choosing lobbying sides: the general data protection regulation of the European Union. *J Publ Policy* 39(4):543–564

- Barik S (2021, August 20) When algorithms dictate your work: Life as a food delivery 'partner'. *Entrackr*. <https://entrackr.com/2021/08/zomato-when-algorithms-dictate-your-work-life-as-a-food-delivery-partner/>
- Bollier D, Helfrich S (2019) *Free, fair, and alive: the insurgent power of the commons*. New Society Publishers
- Chao S (2018) In the shadow of the palm: dispersed ontologies among Marind, West Papua. *Cult Anthropol* 33(4):621–649
- Cooper H, Hirst N (2017, May 4) Silicon Valley tech lobbyists swarm Brussels. *Politico*. <https://www.politico.eu/article/silicon-valley-tech-lobbyists-swarm-brussels/>
- Cormack D, Kukutai T (2022) Indigenous peoples, data, and the coloniality of surveillance. In: *New perspectives in critical data studies: the ambivalences of data power*. Springer International Publishing, pp 121–141
- Cumbers A (2020) *The case for economic democracy*. Polity Press
- De Angelis M (2017) *Omnia Sunt Communia: on the commons and the transformation to Postcapitalism*. Zed Books
- Draper NA (2017) From privacy pragmatist to privacy resigned: challenging narratives of rational choice in digital privacy debates. *Policy Internet* 9(2):232–251
- Dür A (2008) Interest groups in the European Union: how powerful are they? *West*
- Dür A (2019) How interest groups influence public opinion: arguments matter more than the sources. *Eur J Polit Res* 58(2):514–535
- Dür A, Mateo González G (2012) Who lobbies the European Union? National Interest Groups in a multilevel polity. *J Eur Publ Policy* 19(7):969–987
- European Group on Ethics (EGE) (2023) Opinion on democracy in the digital age. Publications Office of the European Union. <https://op.europa.eu/en/publication-detail/-/publication/8b11a1bc-0f21-11ee-b12e-01aa75ed71a1/language-en>; European Politics, 31(6):1212–1230
- Fisher M (2022) *The Chaos Machine: the inside story of how social media rewired our minds and our world*. Hachette UK
- Freund D, Bendel Y (2017) Access all areas: when EU politicians become lobbyists. *Transparency International EU*. <https://transparency.eu/wp-content/uploads/2017/01/Access-all-areas.pdf>
- Goodin RE (2007) Enfranchising all affected interests, and its alternatives. *Philos Public Aff* 35(1):40–68
- Haugen F (2023) *The power of one: blowing the whistle on Facebook*. Hodder and Stoughton
- Hollo T (2022) Welcome to the interregnum. *Meanjin* 81(4):165–171
- Horwitz J, Seetharaman D (2020, May 26) Facebook executives shut down efforts to make the site less divisive. *The Wall Street Journal*. <https://www.wsj.com/articles/facebook-knows-it-encourages-division-top-executives-nixed-solutions-11590507499>
- Kergueno R (2022) The über-lobbyists: how Silicon Valley is changing Brussels lobbying. *Transparency International EU*. <http://transparency.eu/uber-lobbyists/>
- Koenig BA (2014) Have we asked too much of consent? *Hastings Cent Rep* 44(4):33–34
- Kukutai T, Taylor J (2016) *Indigenous data sovereignty: toward an agenda*. ANU Press
- MIT Technology Review Insights (2021) Data fairness: a new social contract for the 21st century economy. *MIT Technology Review Insights*. <https://mittrinsights.s3.amazonaws.com/data-fairness.pdf>
- Ostrom E (2010) Beyond markets and states: polycentric governance of complex economic systems. *Am Econ Rev* 100(3):641–672
- Pasquale F (2017, June 17) From territorial to functional sovereignty: the case of Amazon. *Law and Political Economy Blog*. <https://lpeblog.org/2017/12/06/from-territorial-to-functional-sovereignty-the-case-of-amazon/>
- Prainsack B (2015) Three H's for health—the darker side of big data. *Bioethica Forum* 8(2):40–41
- Prainsack B (2019) Data donation: how to resist the iLeviathan. In: Krutzinna J, Floridi L (eds) *The ethics of medical data donation*. Springer, Berlin, pp 9–22

- Prainsack B (2023) The roots of neglect: towards a sociology of non-imagination. *TECNOSCIENZA: Ital J Sci Technol Stud* 13(2):13–34
- Prainsack B, Forgó N (2022) Why paying individual people for their data is a bad idea. *Nat Med* 28:1989–1991
- Prainsack, B, El-Sayed S, Forgó N, Szoszkiewicz Ł, Baumer P (2022a) Data solidarity: a blueprint for governing health futures. *The Lancet: Digital Health* 4/11: E773–E774.
- Prainsack, B, El-Sayed S, Forgó N, Szoszkiewicz Ł, Baumer P (2022b). Data solidarity – A White Paper. Geneva: The Lancet & Financial Times Commission on Governing Health Futures. Available at: <https://www.governinghealthfutures2030.org/wp-content/uploads/2022/12/DataSolidarity.pdf>
- Purtova N (2022) From knowing by name to targeting: the meaning of identification under the GDPR. *Int Data Priv Law* 12(3):163–183
- Ruohonen J (2019) David and Goliath: Privacy Lobbying in the European Union. arXiv preprint arXiv:1906.01883
- Scott JC (1998) Seeing like a state: how certain schemes to improve the human condition have failed. Yale University Press
- Smith P (2023, July 31) Google tells government how to regulate AI (and who to blame when it goes wrong). *Australian Financial Review*. <https://www.afr.com/technology/google-tells-government-how-to-regulate-ai-and-who-to-blame-when-it-goes-wrong-20230728-p5ds0s>
- van Est R, Gerritsen JBA, with the Assistance of Kool, L (2017) Human rights in the robot age: challenges arising from the use of robotics, artificial intelligence, and virtual and augmented reality. Expert report written for the Committee on Culture, Science, Education and Media of the Parliamentary Assembly of the Council of Europe (PACE). Rathenau Instituut
- Véliz C (2023, July 28) Twitter Post. July 28, 2023, 06:07pm. Last week, the White House announced that it had secured “voluntary commitments” (i.e. pinky promises). Twitter. <https://twitter.com/carissaveliz/status/1684958891358986240?s=46>
- von Arx M (2023) The illusion of immediacy: on the need for human synchronization in data-intensive medicine. *Front Sociol* 8:1120946
- Wagenaar H, Bartels K (2024) Special issue: advancing the commons verse: the political economy of the commons. *Int J Commons* 18(1):337–350
- Wagenaar H, Prainsack B (2021) The return of good government. In: *The pandemic within*. Policy Press, pp 65–83
- Yunkaporta T (2019) *Sand talk: how indigenous thinking can save the world*. Text Publishing

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.



Part II
**Regulatory Compliance: “Here We Must
Run as Fast as We Can, Just to Stay in
Place”**

Regulatory Equivalence in Blockchain Systems: The Role of Public Values and Legitimacy



Primavera De Filippi and Morshed Mannan

Abstract In this chapter, we develop the concept of ‘regulatory equivalence’. As opposed to functional equivalence, used to extend the scope of existing legal frameworks to new technological arrangements, regulatory equivalence refers to the use of technological guarantees to serve the same purpose as traditional legal formalities. This approach goes beyond the use of technology to execute certain legal formalities, but involves analysis of the equivalence between the values undergirding legal rules and the affordances of technological artifacts. This chapter draws on the distinct properties of blockchain-based systems, such as notarization systems, Decentralized Autonomous Organization (DAO)s, and privacy pools, to demonstrate the opportunities and challenges they present to establishing regulatory equivalence. Public actors, however, are hesitant to recognize regulatory equivalence, due to competing perceptions of legitimate governance between participants within these systems and external actors such as regulators. We explore this tension before concluding that this challenge can be overcome through co-regulatory engagement between public authorities and actors within blockchain-based systems, and public authorities more explicitly stating the values they seek to promote within blockchain-based systems.

Funding: This research is funded by the European Research Council under the European Union’s Horizon 2020 Research and Innovation Programme (Grant Agreement No. 865856).

P. De Filippi (✉) · M. Mannan
National Center of Scientific Research (CNRS)/Berkman-Klein Center,
Cambridge, MA, USA

European University Institute (EUI), Fiesole, FI, Italy
e-mail: pdefilippi@cyber.harvard.edu; Morshed.Mannan@eui.eu

1 Introduction

In 2014, one of the co-founders of the Ethereum blockchain, Gavin Wood, argued that a blockchain system is intrinsically “*alegal*” because it “cannot care [...] as to whether its actions might be interpreted as legal or illegal”.¹ Yet, almost a decade later, with the growing adoption of blockchain technology, many actors are now seeking to make blockchain applications regulatorily compliant or, at least, ‘regulatorily equivalent’. Regulatory equivalence builds on the concept of functional equivalence, a term that is well-established in the legal literature.² Functional equivalence allows the establishment of equivalence between two artifacts, one that is already covered within the realm of a legal rule and another that is not (yet) encompassed by it (e.g., written signatures and electronic signatures). Regulatory equivalence goes further by striving to establish the equivalence between the policy objectives of specific legal provisions, and the implications of adopting a particular technological artifact as an alternative way to achieve regulatory objectives.³

This new trend is due in large part to the spectacular bankruptcies (e.g. FTX) and scams (e.g. OneCoin, BitConnect) that have brought actors within the blockchain space in direct confrontation with law enforcement authorities and regulators, but also to a growing interest in pro-actively cultivating the legitimacy of blockchain networks to ensure their long-term sustainability.⁴ To the extent that blockchain applications want to interface with the legal or administrative systems of existing nation states, they must comply with their rules. Regulatory compliance, however, is not always a viable option. Indeed, it can sometimes be challenging to reconcile the demands of the legal system (designed for legacy institutions) with the distinctive properties of blockchain technologies (in terms of e.g. decentralization, transparency, immutability and irreversibility), in ways that do not undermine their basic technological guarantees or their legitimacy within blockchain (‘Web3’) communities.

Regulatory equivalence can be regarded as an alternative way to achieve regulatory compliance, by establishing equivalence between the objective of existing legal formalities and the technological guarantees provided by a new technological artifact. For instance, the COALA DAO Model law discussed in this chapter proposes regulatory equivalence between a variety of requirements for corporate formation and the technological design of specific blockchain-based decentralized autonomous organizations (DAOs). These organizations are not governed by traditional by-laws (i.e. legal and contractual rules) but rather by smart contracts (i.e. technological rules encoded as a computer program on a blockchain network). As such, their rules of operation will be enforced by the underlying blockchain infrastructure, regardless of their compliance with the law.

¹ See Wood (2015).

² See Savin (2019); Finocchiaro and Castellani (2023).

³ See De Filippi et al. (2022b).

⁴ See Thakur (2023), Kerr et al. (2023) and De Filippi et al. (2022c).

At times, technological design can provide equivalent (if not superior) guarantees to traditional legal formalities. For instance, blockchains make it possible to achieve real-time auditing of transactions, thereby reducing the need for third-party audits and period reports. Yet, these technological guarantees may fall short of satisfying some of the motivations that underpin public values, particularly those that are implicit and have not been expressly articulated in the law. Think, for instance, about the public value of transparency, which, in the context of financial transactions, has been explicitly articulated in regulations against money laundering and counter-terrorism financing. Yet, these regulations have been claimed to also be motivated by a more general desire of surveillance by regulatory authorities, with a view not only to prevent illegitimate financial activities, but also to monitor and scrutinize legitimate ones.⁵ Moreover, regulatory equivalence threatens the functions of existing regulatory authorities. For instance, if DAOs were to secure regulatory equivalence with the filing and registration requirements by simply posting their formation and disclosure documents on a secure public forum, at least notionally they could bypass the need for the imprimatur of state authorities like a business registry. This might entail a loss of sources of revenue for the state (e.g., filing and incorporation fees). Consequently, lawmakers and regulators might be reluctant to recognize this type of regulatory equivalence.

At the same time, while blockchain technology continues to develop, these authorities are left facing a regulatory dilemma: curbing the technology's negative effects, without excessively stifling innovation. This has led to regulatory experimentation to support the development of the crypto-industry (e.g., regulatory sandbox regimes, legislation that enables the creation of limited liability DAOs) as well as regulations that curb its perceived risks (e.g., tailoring anti-money laundering and counter-terrorism financing rules to crypto-assets). In the process of drafting these new regulations, it has become necessary for national and global (e.g., the Financial Action Task Force) regulators to articulate the "public values" that they want to promote. This revealed a potential tension between values such as financial privacy and other public values like financial scrutiny and oversight.⁶ The question we are now faced with is whether the technological guarantees afforded by blockchains fulfill public values in a manner that is considered legitimate by society and relevant regulatory bodies?

In theory, if properly implemented, regulatory equivalence can establish greater confidence among all relevant stakeholders than traditional legal formalities and enhance the legitimacy of actors operating in the blockchain space. However, if not properly implemented, reducing the role of publicly accountable trusted intermediaries can lead to accusations that regulatory equivalence creates new publicly unaccountable intermediaries (e.g., miners, token holders, or other blockchain stakeholders), thereby undermining confidence, legitimacy, and legal certainty. Ultimately, to ensure that regulatory equivalence is recognized, we need to

⁵ See Rozenshtein (2018) and Vıcek (2008).

⁶ See Cherednychenko (2021).

guarantee that the underlying governance is robust, and that there is enough accountability at the governance level.

This chapter starts by elucidating the principles of functional equivalence (Sect. 2.1) and regulatory equivalence (Sect. 2.2) by providing a series of examples to illustrate the distinction between the two. While the former has largely been recognized by regulators as a useful tool to extend the scope of regulations to novel technologies without having to undertake extensive legislative reforms, the latter is a more contentious principle whereby regulation can recognize alternative ways to achieve regulatory compliance by leveraging the specific characteristics of novel technologies. The chapter continues by presenting the unique opportunities (Sect. 3.1) and challenges (Sect. 3.2) that regulators must face in the process of establishing regulatory equivalence in the context of blockchain systems. Drawing from selected examples (e.g. notarization systems, DAO Model Law, and privacy pools), it is shown that while there are tangible benefits in recognizing the use of technological guarantees as an alternative means for blockchain systems to achieve regulatory compliance, such approach is not devoid of challenges, in particular when the technological guarantees fulfill some, but not all, of public values—whether these are explicitly or implicitly enshrined within the law.

Finally, the chapter tackles the issue of legitimacy in the blockchain systems, and the need to resolve the tension between endogenous legitimacy (i.e. how the operations of these systems are perceived by the system's participants) and exogenous legitimacy (i.e. how these systems are perceived by external actors, including regulators and society at large). On the one hand, regulators will only recognize regulatory equivalence to systems that are perceived as legitimate by societal actors; on the other hand, blockchain systems will only try to implement technological guarantees to implement specific regulatory objectives which are perceived as legitimate by the system's participants. The chapter concludes with the claim that, in order for regulatory equivalence to be properly implemented by blockchain actors and effectively recognized by regulators, it is crucial for public authorities to more explicitly state the values they are ultimately seeking to promote, as well as the activities they are seeking to repress. Only then does it become possible for blockchain developers and practitioners to properly understand the policy objectives that the systems they are building need to satisfy in order to qualify for regulatory equivalence.

2 Functional Equivalence and Regulatory Equivalence

In this section, we will first define our usage of specific terms: *functional equivalence* and *regulatory equivalence*. Each concept is illustrated with examples and respectively distinguished from *technological neutrality* and *regulation by design*. We show how regulatory equivalence represents an evolution of functional equivalence, prefiguring a future regulatory environment in which public policies and technological innovation coexist more harmoniously than at present.

2.1 *Functional Equivalence*

We first turn to the notion of *functional equivalence*. Certain artifacts, such as paper documents, hand-written signatures, and physical seals of wax, rubber or steel, have for centuries been used for the execution of contracts and deeds. Across the various legal systems of the world, these artifacts are recognized as having legal effects and are required for the valid execution of a wide range of transactions. For instance, the law traditionally prescribes that the creation and disposal of interests in land (e.g., the sale of land) and the preparation of wills must be in writing.⁷ While oral contracts are common, in written contracts a signature by the contracting parties (and in some cases witnesses)⁸ is necessary for contracts to be finalized and to enter into force, unless agreed otherwise. In the case of agreements that have a public dimension or involve foreign transactions, such as a land transfer or marriage, a third party—such as a land registry,⁹ notary, and/or (other) state authority—must also affix their seal(s) to the document. Dalgetty,¹⁰ for instance, observes that in England and Wales every “practicing notary has their own unique seal embossed on a document which can only be used by him or her when completing a notarial act. This must be in the form of a ‘tamper proof’ document, and is usually sewn with a ribbon and sealed”.¹¹ In short, these *physical artifacts* have significant implications for the validity of legal documents and are designed with certain public interests in mind, such as transaction security and fraud protection.

With the development of electronic media and commerce in the latter half of the twentieth century, new *digital artifacts* have emerged that can serve the same functions as analogue physical artifacts while protecting public interests such as transaction security and fraud protection, provided that certain legal and technical requirements are met. For instance, the introduction to the UNCITRAL Model Law for Electronic Commerce (1996) mentions that electronic records can provide at least the same degree of security as a paper document.¹² In addition, electronic records can be widely legible, tamper-proof, reproducible, authenticated by a signature, and submitted to public authorities. They can even fulfill some of these functions more expeditiously and reliably than ordinary paper documents. In the following years, the use of electronic documents has become commonplace. For instance, under English law, the electronic conveyance of land has become possible, and an electronic document can be legally deemed to be a land deed.¹³ Similarly, under the UNCITRAL Model Law, a ‘data message’ can be deemed to be equivalent

⁷Law of Property Act 1925 (15 & 16 Geo. 5. c. 20), section 53(1)(a), Wills Act 1837 (c. 26 7 Will 4 and 1 Vict) and section 9(1)(a) (England & Wales).

⁸Wills Act 1837 (c. 26 7 Will 4 and 1 Vict) and section 9(1)(c)-(d) (England & Wales).

⁹Land Registration Act 2002 (2002 c. 9) and Sch. 7 para 5 (England & Wales).

¹⁰See Dalgetty (2011).

¹¹Dalgetty (2011), p. 325.

¹²See UNCITRAL (1996).

¹³Land Registration Act 2002, section 91 (England & Wales).

to certain signatures, if the following conditions are met: (1) there is a method for identifying the person sending the data message, (2) there is a method for showing that a person approved the information in the message, and (3) the method is reliable under the circumstances.¹⁴ UNCITRAL acknowledged that there are different types of signature requirements (e.g., additional seals, or witness signatures) for various transactions and further functional equivalents may need to be developed to accommodate them.¹⁵ Indeed, various ‘levels’ of electronic signatures and electronic seals are recognized under EU law. Following the passage of the EU’s eIDAS Regulation, all Member States recognized the validity of three types of electronic seals as being equivalent to comparable physical seals.¹⁶ As such, they have been recognized by the law as having the same function or effects as the corresponding physical artifacts—thereby acknowledging their functional equivalence.

In short, functional equivalence is the legal recognition that an artifact not encompassed by the law can fulfill the same functions as an artifact that is already encompassed by the law, thereby extending the variety of artifacts encompassed by the law. As explained by UNCITRAL, a functional equivalence analysis involves first analyzing the functions that a paper-based method serves before establishing whether another technology can fulfill the same functions. If such equivalence can be determined, a new rule can be drafted that encompasses this artifact.¹⁷

Blockchain artifacts can potentially be functionally equivalent to earlier digital and physical artifacts. For instance, jurisdictions like Delaware and Vermont in the United States and Poland in the European Union already permit companies to maintain their share registers on blockchain-based systems,¹⁸ treating them as being functionally equivalent to other non-distributed types of share registers subject to limited statutory requirements. However, a functional equivalence analysis may have unintended, even adverse, consequences for persons and organizations developing blockchain artifacts. Take, for example, the classification of tokens issued by blockchain protocols as (unregistered) securities in the United States. At first glance, the application of an investment contract classification test (i.e., the *Howey* test) to crypto tokens, which were originally applied to interests in orange groves, can be seen as an example of functional equivalence. Under this perspective, the crypto tokens recorded on a blockchain network are deemed to be equivalent to the paper documents that traditionally comprise an investment contract.¹⁹ However, this is not an example of functional equivalence but rather an example of the law’s use of analogical reasoning and adherence to neutrality of technology.

¹⁴ UNCITRAL Model Law for Electronic Commerce, art. 7.

¹⁵ UNCITRAL Model Law for Electronic Commerce, p. 38.

¹⁶ See Mason (2018).

¹⁷ Harvey (2017), p. 60.

¹⁸ See Mazgaj (2021).

¹⁹ See Henning (2018) and Collomb et al. (2019).

While functional equivalence has occasionally been equated with neutrality of technology, they are distinct concepts.²⁰ Neutrality of technology is broadly concerned with ensuring that the *effects* of online and offline acts are treated in the same way. The law is thereby “indifferent” to the technology used.²¹ For instance, the act of defamation should be subject to the same legal treatment irrespective of whether the defamatory act is committed online on a chat forum or offline in a print newspaper. In the US, under the *Howey* test, it appears that in determining the existence of an investment contract, the technological medium is irrelevant. However, the results of such an assessment are not self-evident in the European Union, as EU law does not have an equivalent of the *Howey* test. The functional assessment of whether crypto assets are ‘transferable securities’ under the EU’s Markets in Financial Instruments Directive (MiFID II), though quite similar, is not identical to the *Howey* test.²² This may lead to a crypto asset being classified differently in the EU compared to the United States.

The application of analogical reasoning and neutrality of technology can be seen in recent US cases concerning the classification of unregistered decentralized autonomous organizations (DAOs). As such DAOs have not made a formal entity choice and there is no legal framework specifically tailored to unregistered DAOs, courts have drawn legal analogies with other legal forms, most notably general partnerships.²³ Again, the focus of courts thus far has been on the substantive factual relationship between human parties in the DAO, rather than the technologies that are used for distributed coordination and/or automation of operations and governance processes.

In both instances, laws and regulations that were designed for centralized systems are being applied to decentralized systems. This can have unintended consequences. For instance, deeming unregistered DAOs to be general partnerships can lead to the joint and several liability of all members of the DAO,²⁴ leaving those members in the jurisdiction of the court that made such determination especially vulnerable. This creates uncertainty, places considerable financial and non-financial risk on DAOs members, and stymies technological innovation. Arguably, such an approach also anchors DAOs and crypto tokens to “outdated characteristics and properties” rather than recognizing their emergent and innovative features.²⁵

Some technologically progressive jurisdictions like England and Wales have come to appreciate these difficulties and have tailored solutions that are suited to decentralized systems. An example of this is the service of notice of legal proceedings and other legal documents and orders using non-fungible tokens (NFTs). English courts have long permitted a variety of mediums for valid service, but

²⁰ See Veerpalu (2019).

²¹ Reed (2010), p. 249.

²² See Ferrari (2020).

²³ *Sarcuni v. bZx Dao*, 22-cv-618-LAB-DEB (S.D. Cal. Mar. 27, 2023).

²⁴ *Sarcuni v. bZx Dao*, 22-cv-618-LAB-DEB (S.D. Cal. Mar. 27, 2023).

²⁵ Harvey (2017), p. 63.

recognizing the challenges that the pseudonymity and non-hierarchical nature of blockchain networks present, courts have recently begun deeming the sending of an NFT containing such documents and orders to blockchain addresses to be valid service.²⁶ This has been possible as the civil procedure rules of English courts already recognized electronic methods of service and, when presented with a sufficiently novel and significant case, this was capaciously interpreted to include NFT airdrops of legal documents. In other words, this new blockchain artifact was deemed to be functionally equivalent to earlier digital artifacts such as email.

2.2 Regulatory Equivalence

Blockchain technology introduces a unique set of challenges when it comes to complying with existing legal provisions crafted for centralized institutions.²⁷ In particular, several difficulties emerge due to the need to accommodate the demands of the legal system with the fundamental properties of blockchain technology.²⁸ Indeed, traditional legal frameworks, established over decades to regulate centralized operators and their platforms, are generally ill-equipped to account for the decentralized and disintermediated nature of blockchain networks.²⁹

Blockchain networks operate on a worldwide scale, with transactions occurring across borders without reliance on centralized operators. The absence of a central governing body creates potential jurisdictional challenges, making it difficult to determine which jurisdiction has authority in case of legal disputes.³⁰ Traditional legal frameworks, designed around national boundaries, may therefore struggle to assert authority over decentralized blockchain platforms with an international user base.³¹

Besides, legal systems traditionally presuppose a set of identifiable entities that can be held accountable for the services they provide to their users.³² Yet, blockchain networks operate in a peer-to-peer manner, as transactions are processed and verified by network participants rather than by a central authority.³³ The pseudonymity inherent in these networks also challenges traditional legal norms, making it difficult to identify parties in case of disputes or wrongdoing.³⁴ Striking a balance

²⁶ *D'Aloia v Person Unknown & Ors* [2022] EWHC 1723 (Ch), *Osbourne v Persons Unknown Category A & Ors* [2023] EWHC 39 (KB), *Osbourne v Persons Unknown & Ors* [2023] EWHC 340 (KB).

²⁷ See De Filippi et al. (2022b).

²⁸ See Yeung (2019).

²⁹ See Magnuson (2018).

³⁰ See De Filippi and Wright (2018).

³¹ See Herian (2018).

³² See Taddeo and Floridi (2016).

³³ See Pereira et al. (2019).

³⁴ See Yeoh (2017).

between privacy expectations of blockchain users and the identification requirements of legal authorities is a delicate task that requires careful consideration.³⁵

Integral to many blockchain applications are smart contracts, i.e., software used to facilitate social and economic transactions through deterministic computation.³⁶ These applications challenge the enforcement and adjudication of contractual disputes because,³⁷ independently of the decision of a judge, no one has the power to unilaterally influence, or even stop the execution of a smart contract.³⁸ This stands in contrast with conventional legal doctrines of contractual interpretation which rely on the human interpretation of natural language to define the nature and scope of any contractual relationship.³⁹ This adds a new layer of complexity to regulatory efforts seeking to govern the operations of blockchain systems.⁴⁰

At the same time, blockchain technology brings new opportunities for regulatory compliance.⁴¹ Traditional legal systems often emphasize the need for more transparency and traceability in centralized institutions. These features, which are inherent in many blockchain networks, provide a significant degree of accountability and verifiability.⁴² As such, they could potentially contribute to, rather than jeopardize, the ability of blockchain systems to comply with existing regulatory requirements.⁴³ Hence, while traditional legal frameworks struggle at adapting to the new challenges and opportunities offered by blockchain technology, blockchain practitioners can fill up the regulatory gap, through the introduction of technological guarantees intended to fulfill specific policy objectives. The applications deployed on blockchain networks can evolve rapidly, driven by technological advancements and community-driven updates that by far outpace the legislative process of traditional legal frameworks.⁴⁴

The increased adoption of blockchain technology in various sectors has enabled new forms of experimentation,⁴⁵ using blockchain technology as a regulatory technology in ways that differ from traditional notions of regulatory compliance centered solely on the fulfillment of existing legal formalities. This is what we refer to as ‘regulatory equivalence’.⁴⁶ Rather than seeking strict adherence to established legal formalities, regulatory equivalence seeks alignment with the overarching policy objectives embedded within specific legal provisions. This is grounded on the

³⁵ See Renwick and Gleasure (2021).

³⁶ See De Filippi et al. (2020).

³⁷ See Werbach (2018).

³⁸ See Raskin (2016).

³⁹ See Temte (2019).

⁴⁰ See Werbach and Cornell (2017).

⁴¹ See De Filippi and Hassan (2018).

⁴² See Shin and Hwang (2020).

⁴³ See De Filippi et al. (2022a).

⁴⁴ See De Filippi (2021).

⁴⁵ See Al-Jaroodi and Mohamed (2019).

⁴⁶ See De Filippi et al. (2022b).

fact that, while the distinctive properties of blockchain technology may not easily conform to existing legal norms, they might nonetheless be instrumental to achieving identical policy goals—albeit via different means.⁴⁷ The emphasis on policy objectives allows for a more flexible interpretation of regulatory requirements, enabling regulators to evaluate the effectiveness of blockchain technology in fulfilling overarching policy goals through technological guarantees,⁴⁸ without strictly adhering to predefined legal formalities.

Regulatory equivalence goes beyond acknowledging technological guarantees as mere substitutes for legal formalities;⁴⁹ it endeavors to establish an actual and meaningful equivalence between the purpose of the legal formalities and implications of these technological guarantees. This requires a comprehensive evaluation of how the adoption of these technological guarantees aligns with the intended policy outcomes of existing legal provisions.⁵⁰ For instance, the transparency and tamper-resistance of a blockchain introduce new opportunities for real-time auditing,⁵¹ which can be regarded as an effective alternative to traditional third-party audits,⁵² potentially streamlining compliance processes.⁵³ The adoption of technological guarantees as an alternative means to comply with specific policy objectives and regulations (instead of complying with traditional legal formalities) forms the basis of regulatory equivalence.⁵⁴ Indeed, identifying these technological guarantees makes it possible to leverage the intrinsic capabilities of blockchain technology without compromising the integrity of traditional regulatory frameworks.⁵⁵

As such, regulatory equivalence can be regarded as a *proactive* response to the inherent challenges faced by blockchain applications seeking integration with traditional legal and administrative systems. Regulatory equivalence stands in contrast to other concepts, such as *regulation by design* and *compliance by design*, which involve the *ex ante* implementation of specific legal requirements in technological design.⁵⁶ Rather than viewing compliance as a rigid adherence to predefined legal steps, regulatory equivalence encourages a forward-looking approach that explores innovative ways to achieve regulatory objectives in ways that simply could not have been foreseen at the time in which legislation was drafted.

Overall, this represents a shift towards a more dynamic and policy-oriented approach to regulation, fostering an environment where technological innovation

⁴⁷ See Spouse and Green (2019).

⁴⁸ See Bodó and De Filippi (2022).

⁴⁹ See Micheler and Whaley (2020).

⁵⁰ See Collomb et al. (2018).

⁵¹ See Abreu et al. (2018).

⁵² See Bonsón and Bednárová (2019).

⁵³ See Dai and Vasarhelyi (2017).

⁵⁴ See Cooper and Lodder (2023).

⁵⁵ See Zwitter and Hazenberg (2020).

⁵⁶ Almada (2023) and Rogers (2023), p. 143.

and public policy can coexist harmoniously.⁵⁷ Yet, the pursuit of regulatory equivalence necessitates an intricate understanding of the relationship between technology and established legal frameworks. It underscores the need for greater legal flexibility and regulatory adaptability in the face of rapid technological advancements, recognizing that new technologies may offer alternative means to comply with policy goals. By focusing on policy objectives, regulators could assess the compatibility of blockchain technology with existing legal frameworks, ensuring that the transformative potential of these emergent technologies are harnessed to support, rather than violate the law.⁵⁸ Failure to address these dynamics may not only impede the legal compliance of blockchain applications but also risk stifling the innovative potential of this technology.

3 Opportunities and Challenges of Regulatory Equivalence for Blockchain

Having explained the concepts of functional and regulatory equivalence, this section presents the main opportunities and challenges of recognizing regulatory equivalence, using examples of blockchain-based notarization processes, DAOs, and privacy pools. The section concludes by considering the main obstacles to regulatory equivalence being recognized by regulators.

3.1 Opportunities

The pursuit of regulatory equivalence presents a series of opportunities for both regulators and regulated entities alike, who might benefit from the identification of new means of achieving public policy and legal objectives.⁵⁹ Yet, it also introduces a host of challenges stemming from the need to reconcile overarching policy objectives with the unique guarantees offered by emergent technologies.⁶⁰ In the realm of blockchain technology, achieving regulatory equivalence requires understanding the technological properties of blockchain systems, in terms of decentralization and disintermediation. One key challenge lies in the reconciliation of existing policy objectives with the requirements of pseudonymity, immutability, and tamper-resistance endorsed by many of these systems.⁶¹ This requires finding the right balance to ensure that such reconciliation does not compromise the fundamental

⁵⁷ See Collomb et al. (2019).

⁵⁸ See Allena (2022).

⁵⁹ See Egan (2002).

⁶⁰ See Collomb et al. (2019).

⁶¹ See De Filippi et al. (2022b).

principles of the legal system, nor the technological guarantees of blockchain technology.

Blockchain technology offers a series of technological guarantees that can potentially fulfill functions associated with traditional legal formalities. One interesting example of this is the undertaking of *notarization processes with blockchain systems*, where decentralized and cryptographic mechanisms could replace the need for a centralized trusted authority.⁶² Regulatory equivalence in this context requires establishing a comparable level of confidence, predictability, and security through technological means. This hinges on a few technological guarantees provided by blockchain technology. First, each block in a blockchain contains a timestamp and a cryptographic hash that links to the previous block, thereby producing an immutable and chronological chain of transactions, with cryptographic primitives ensuring the authenticity and integrity of all data recorded therein.⁶³ Recording the hash of a digital document into a blockchain transaction thus provides tamper-resistant proof that such a document or transaction existed at a specific point in time.⁶⁴ Second, because users employ cryptographic signatures to validate their transactions, the mere act of recording the hash of digital files into a blockchain transaction constitutes a secure and non-repudiable means of validating the source or provenance of these digital documents.⁶⁵

Third, the transparent nature of a public blockchain allows all participants to view and verify the recorded transactions. This transparency ensures that the information recorded on the blockchain always remains accessible to relevant parties, creating a shared and publicly verifiable record. This transparency can be regarded as equivalent to public record-keeping functions traditionally associated with notary services.⁶⁶ Combined with the technological guarantees provided by blockchain technology (in terms of cryptographic security, immutability, tamper-resistance, non-repudiability, as well as public timestamping), a blockchain can be regarded as achieving equivalent functions as traditional notary services, concerning document certification and fraud prevention, with at least an equivalent level of assurance and legal certainty.⁶⁷

Blockchain technology also offers a set of technological guarantees that can fulfill functions analogous to legal formalities required for traditional corporate entities. This can be illustrated through the example of DAOs.⁶⁸ As discussed above, unregistered DAOs do not enjoy the same benefits as traditional corporate entities and are generally assimilated to a simple form of general partnership.⁶⁹ The

⁶² See Nasti (2021).

⁶³ See De Filippi and Wright (2018).

⁶⁴ See Arredondo (2018).

⁶⁵ See Putz et al. (2019).

⁶⁶ See Lemieux et al. (2019).

⁶⁷ See Ulloa and Gallegos (2021).

⁶⁸ See Hassan and De Filippi (2021).

⁶⁹ See Morrison et al. (2020) and Guillaume (2023).

challenge here is to establish regulatory equivalence by aligning the policy objectives of legal provisions governing corporate formations with the unique technological guarantees of blockchain-based DAOs. The *DAO Model Law*⁷⁰ attempts to address this problem with regulatory equivalence, claiming that the transparency and immutability intrinsic to blockchain technology can fulfill, under specific conditions, the registration formalities that traditional corporations are legally required to abide by.⁷¹ First, a DAO is operated via smart contracts, i.e., self-executing code deployed on a blockchain. Once deployed, these contracts are executed by the blockchain network as a whole, meaning they cannot be unilaterally altered or tampered with.⁷² These smart contracts are therefore, to some extent, akin to legal documents associated with traditional corporate formalities, such as articles of incorporation, that do not have to be updated often.⁷³

Second, DAOs often use tokens to represent ownership and voting rights, where token holders participate in decision-making processes in proportion to their token holdings.⁷⁴ This mirrors the governance structures inherent in traditional corporate entities, where decisions are often made through shareholder voting or board resolutions.⁷⁵ Thus, token-based ownership and voting mechanisms can be regarded as a technological evolution of the paper and electronic share system in traditional corporations. Third, all transactions and decisions within DAOs are recorded on a public blockchain, acting as a transparent and publicly accessible record. This aligns with the policy objectives of corporate law, concerning disclosure requirements. The public accessibility of DAO records guarantees that stakeholders have relevant insights into the governance processes of a DAO, analogous to the transparency requirements in traditional corporate entities.⁷⁶ Finally, the immutability of a blockchain ledger functions as an additional technological guarantee to achieve regulatory equivalence, by maintaining a clear record of governance actions, that cannot be unilaterally altered or manipulated.⁷⁷

Privacy Pools provide another interesting example of an attempt at achieving regulatory equivalence. Privacy Pools represent an innovative approach that leverages cryptography and the technological guarantees of blockchain technology to strike a balance between maintaining user privacy in financial transactions while

⁷⁰ The DAO Model Law is an initiative of COALA (Coalition of Automated Legal Applications) that outlines a blueprint for the regulation of DAOs, offering a coherent and adaptable framework that jurisdictions can reference in their efforts to accommodate this novel form of organization. It delineates the extent to which some DAO features can be held to be functionally equivalent, and therefore compliant, with various legal requirements, while others might be regarded as regulatory equivalent in that they achieve the same policy objectives. See Choi et al. (2021).

⁷¹ Choi et al. (2021).

⁷² See Sulkowski (2019).

⁷³ See Reyes (2019).

⁷⁴ See Mannan (2023).

⁷⁵ See Mahmoud (2022).

⁷⁶ See Borgogno (2022).

⁷⁷ See Sims (2019).

fulfilling analogous policy objectives as those underpinning the transparency requirements imposed by law concerning financial transactions.⁷⁸ Privacy Pools implement confidential transaction protocols that obfuscate the link between incoming and outgoing transactions. These protocols typically rely on ring signatures and anonymity sets, making it challenging to attribute transactions to specific individuals within the pool.⁷⁹ Yet, they also incorporate selective mechanisms that allow users to control the inclusion or exclusion of specific wallets in the pool.⁸⁰ This functionality is particularly relevant in the context of meeting the regulatory objectives of laws, for instance focused in Counter-Terrorism Financing (CTF), where users may seek to prevent association with wallets linked to criminal activities. Some privacy pools may offer users the ability to create and manage a whitelist of specific wallets that they trust or consider as legitimate counterparts. Others may offer a mechanism to exclude wallets associated with illicit or criminal activities by maintaining a blacklist.⁸¹ This feature enables users to proactively distance themselves from wallets linked to unlawful activities, aligning with regulatory objectives and contributing to a more secure and trustworthy environment within the privacy pool. Moreover, Privacy Pools can also incorporate mechanisms allowing for the selective disclosure of transaction details to specific entities (e.g., regulatory authorities), if desirable by the parties who took part in the transaction.⁸² Instead of a binary choice between complete transparency (as in the case of most public blockchains) or complete anonymity (as in the case of traditional cryptocurrency tumblers), selective disclosure mechanisms enable regulated entities to disclose specific transaction information, while still preserving the overall privacy of transactions.

3.2 Challenges

Despite their benefits, none of the technological artifacts described above have, thus far, benefited from regulatory equivalence. The lack of recognition for regulatory equivalence in these three examples can be attributed to multifaceted challenges that extend beyond the technological domain. Even if—technically speaking—blockchain solutions do implement some of the technological guarantees necessary to address a particular policy objective, a variety of factors have to be taken into account when establishing regulatory equivalence, including, but not limited to legal, economic, and sociopolitical considerations. Thus, even though the pursuit of regulatory equivalence presents opportunities for both regulators and regulated entities eager to promote and facilitate regulatory compliance, new challenges arise

⁷⁸ See Buterin et al. (2023).

⁷⁹ See Cirkovic (2022).

⁸⁰ See Beal and Fisch (2023).

⁸¹ See Wang et al. (2023).

⁸² See Zhou et al. (2023).

from the difficult reconciliation of overarching policy objectives with the distinctive properties of blockchain technology.

Despite early attempts by countries like Estonia to implement new forms of blockchain-based notarization within its e-residency program,⁸³ embracing regulatory equivalence in the context of public and permissionless blockchain systems is limited by three important challenges. First, while blockchains can provide tamper-resistant proofs and non-repudiable means of validating digital documents, challenges persist in *verifying the identity of pseudonymous parties* involved in transactions and ensuring that the parties have, indeed, properly understood the terms of the transaction.⁸⁴ Second, one must ensure that the blockchain network on which notarization services are provided is sufficiently secure not to compromise the robustness of technological guarantees, e.g., as a result of a 51% attack.⁸⁵ Third, and perhaps most importantly, recognizing regulatory equivalence for notarization services through disintermediated blockchain-based systems could be perceived as detrimental to public institutions due to the potential loss of revenue generated from traditional notary fees and services.⁸⁶

When it comes to the legal recognition of DAOs, the challenge lies in aligning the policy objectives of legal provisions governing corporate formations with the unique technological guarantees of blockchain-based systems. The DAO Model Law is an attempt to establish regulatory equivalence by showing that the deployment of smart contracts governing a DAO on a publicly accessible, transparent and immutable system can be assimilated to traditional registration formalities.⁸⁷ Yet, two important challenges remain. First, it is crucial to precisely stipulate the *criteria for a DAO to qualify for legal recognition* and how to ensure that these conditions remain true over time, even when the membership of the DAO—and therefore its governance—has changed.⁸⁸ Second, *foregoing registration may be seen as disadvantageous for public institutions as it may result in a reduction of revenue* typically generated through state registration processes for corporate entities, including fees for filing and incorporation.

Privacy Pools also encounter challenges in achieving regulatory equivalence due to the delicate balance between *user privacy and transparency requirements*. While privacy pools leverage cryptographic techniques to obfuscate transaction details while maintaining a certain degree of control over who can transact with whom,⁸⁹ challenges emerge in determining how these privacy features can coexist with the information disclosure needed for legal compliance, such as anti-money laundering

⁸³ See Sullivan and Burger (2017).

⁸⁴ See Koos (2023).

⁸⁵ See Ametrano (2016).

⁸⁶ See Dias Menezes (2023).

⁸⁷ See Choi et al. (2021).

⁸⁸ See Tse (2020) and Gonzalez (2023).

⁸⁹ See Buterin et al. (2023).

laws.⁹⁰ The selective disclosure mechanisms must strike the right balance to satisfy both user privacy concerns and regulatory transparency mandates.

In sum, the reluctance of regulators to recognize regulatory equivalence in the field of blockchain technology can be attributed to various factors. One of the primary concerns is the perceived threats to the public institutions responsible for upholding and enforcing laws. *The use of blockchain technology as a new mechanism for regulatory compliance may diminish the supervisory powers traditionally held by regulatory authorities and legal intermediaries.*⁹¹ Indeed, the decentralization and disintermediation inherent in many blockchain systems, along with their self-executing and decentralized governance structures, may alter the dynamics of regulatory control, shifting decision-making power from centralized authorities to a distributed network of non-identified participants.⁹² This challenges the traditional hierarchical structure of law enforcement,⁹³ contributing to the reluctance of regulators to recognize the regulatory equivalence of blockchain artifacts to the policy objectives of existing legal frameworks.

A further concern with the recognition of blockchain technology as a means of regulatory compliance is the potential *loss of revenue for public administrations*. Existing regulatory processes often involve fees, taxes, and other revenue sources collected by public institutions. If blockchain technology can streamline or circumvent these processes, it could generate a direct financial loss for public services. This may further contribute to the hesitancy of regulators to acknowledge blockchain-based regulatory equivalence.

Finally, this hesitancy might also stem from the difficulty of ensuring that blockchain-based systems correspond to and uphold liberal democratic values.⁹⁴ The decentralized and pseudonymous nature of blockchain technology, while offering unique advantages, *may conflict with principles of fairness, equality, accountability, and public scrutiny inherent in liberal democracies.*⁹⁵ Regulators may be concerned that recognizing regulatory equivalence in blockchain systems could inadvertently lead to the creation of new, non-publicly accountable intermediaries within the blockchain ecosystem—such as miners, validators, token holders, or other relevant stakeholders. We discuss this further in the section below, by expanding on the *legitimacy* of blockchain-based guarantees in the context of regulatory equivalence.

These challenges show that establishing regulatory equivalence in the blockchain landscape requires striking a delicate balance. Regulators must carefully weigh the benefits of a blockchain's technological guarantees against the potential threats to established authorities, democratic values, and existing revenue models.

⁹⁰ See Silva Ramalho and Igreja Matos (2021).

⁹¹ See Meijer and Ubacht (2018).

⁹² See Atzori (2015).

⁹³ See Miscione and Kavanagh (2015).

⁹⁴ See De Filippi (2019).

⁹⁵ See Cossar et al. (2022).

This requires a nuanced understanding of the interplay between the explicit policy objectives recognized by the law, and the more implicit political and economic considerations underpinning these laws. Those advocating for regulatory equivalence must therefore engage in careful consideration to demonstrate how a blockchain system is not only an analogous (or advantageous) alternative to existing legal formalities but also provides economic and political equivalents to the functions traditionally served by these legal formalities.

4 Public Values and Legitimacy of Blockchain Technology

In the preceding section we explained the opportunities and challenges of recognizing the technological guarantees of blockchain systems as being regulatorily equivalent to the objective enshrined in existing laws and regulations. In addition to the complexities inherent in understanding the technological properties and consequences of blockchain artifacts, public authorities may also be concerned that recognition of regulatory equivalence for these artifacts would not be seen as *legitimate*, to the extent that doing so would be contrary to public values. In this section, we evaluate the extent to which the technological guarantees provided by blockchain-based systems can conform to public values and thereby be considered as legitimate means of achieving specific regulatory objectives. Subsequently, we consider reforms that, in the spirit of *co-creation* and *co-regulation*,⁹⁶ would enhance the legitimacy of these technological guarantees and consequently enable the recognition of regulatory equivalence in blockchain systems.

Public values are at the heart of several definitions of the concept of *legitimacy*. In Max Weber's famous articulation of the concept, the authority of a political system is considered to be legitimate if the exercise of power is congruent with the values of the people that are being governed by the system.⁹⁷ Similarly, Talcott Parsons acknowledged that consistency between the values of governors and the governed can help generate legitimacy within a political system.⁹⁸ As such, the legitimacy of power arises from powerholders exercising power in conformity with what is perceived as being proper or desirable in a broader, socially constructed value system.⁹⁹ Margaret Levi expands on this further, by adding that beyond the governors and the governed sharing higher-order values and moral principles, there need to be mechanisms for the governed to have a voice in governance.¹⁰⁰ Naturally, public values differ across societies. What may be deemed to be legitimate in one society, may not be deemed to be legitimate in another due to differing value

⁹⁶ See Gertz and Ozkaramanli (2024).

⁹⁷ See Weber (1964).

⁹⁸ See Parsons (1960).

⁹⁹ Suchman (1995) p. 574.

¹⁰⁰ See Levi (2019).

systems. Values that are deemed to be important in liberal democratic political systems, may not always be considered appropriate or desirable in the context of blockchain systems ('Web3 ecosystems').¹⁰¹ Relatedly, there are dramatic differences in the ways in which various systems enable input into governance processes, and the weight they give to different forms of input.

The topic of legitimacy has long been discussed in political science and legal theory, but it has also generated considerable interest in the Web3 ecosystem following the publication of Ethereum co-founder Vitalik Buterin's blog post 'The Most Important Scarce Resource in Blockchain: Legitimacy' in 2021.¹⁰² As coercion is more difficult in public permissionless blockchain-based systems relative to traditional political systems (given the ease with which it is possible to exit these systems and the opportunities there are to contribute to the development, maintenance and governance of these open-source technologies), the conditions for the existence of legitimacy also differ. For Buterin, legitimacy in the Web3 ecosystem emerges from a community's acceptance of a blockchain system, which can be indicated by the continued usage of the system and the capital mobilized into the system. Indeed, in an ecosystem where participation is voluntary, legitimacy acts as a social glue that keeps users loyal to the system and helps generate the network effects necessary for the system to be viable.¹⁰³ There are various ways in which legitimacy can be established in a blockchain system—from the exercise of force by a charismatic authoritarian ruler to broad-based participation—but their appropriateness will be determined by the values of that system. "Brute force", for instance, is unlikely to be desirable, or even feasible, in a Web3 ecosystem that values "freedom", "decentralization", and credible neutrality¹⁰⁴

So far, we have discussed legitimacy in Web3 ecosystems in terms of endogenous legitimacy, i.e., the perceptions of communities within those ecosystems. However, *exogenous legitimacy*, i.e., how Web3 ecosystems are perceived by external actors, such as the state, has not yet been addressed. The perception of external actors turns on the type of legitimacy they are concerned with. Dimitropoulos,¹⁰⁵ for instance, explains that the use of blockchain technologies for humanitarian purposes by international organizations may be perceived as being legitimate by the public from a substantive and output-oriented view, to the extent that they can improve the efficiency of aid disbursement processes and promote certain ideals such as upholding human rights. However, this use may be perceived as being

¹⁰¹ See De Filippi et al. (2022a).

¹⁰² See Buterin (2021).

¹⁰³ See De Filippi et al. (2022a) and Rosati et al. (2021).

¹⁰⁴ Buterin explains that a mechanism is credibly neutral if "just by looking at the mechanism's design, it is easy to see that the mechanism does not discriminate for or against any specific people. The mechanism treats everyone fairly, to the extent that it's possible to treat people fairly in a world where everyone's capabilities and needs are so different. "Anyone who mines a block gets 2 ETH" is credibly neutral, "Bob gets 1000 coins because we know he's written a lot of code and we should reward him" is not."" See Buterin (2020) and Buterin (2021).

¹⁰⁵ See Dimitropoulos (2022).

illegitimate by the public in terms of input and procedural legitimacy, given the potential lack of meaningful input from vulnerable groups and public actors when designing the system.¹⁰⁶ This shortcoming is increasingly present in humanitarian projects where international organizations have begun using private, permissioned blockchains—thereby inhibiting who can contribute to and decide on their development.

However, even if humanitarian applications were built on public, permissionless blockchains, there may still be concerns about the extent to which the views of vulnerable groups and public actors are given consideration, relative to those of private actors in Web3 ecosystems, such as software developers and (at least indirectly) miners and validators.¹⁰⁷ Reinsberg concedes that blockchain systems could entrench existing power asymmetries but is more optimistic than Dimitropoulos about their input and procedural legitimacy, especially when it comes to global climate governance.¹⁰⁸ According to him, the use of blockchain-based systems could not only broaden the number of people who can read or record blockchain transactions, but also diversify the sources of data to directly include local, marginalized communities affected by climate change (e.g., through the use of oracles). This can contribute to more democratic legitimacy and accountability of global climate governance, and more broadly, prefigure a “more liberal” global governance system.¹⁰⁹

Accordingly, the extent to which the use of blockchain-based systems by organizations is viewed as exogenously legitimate clearly depends on multiple factors. The examples mentioned above are, at least ostensibly, designed for beneficial social purposes, thereby benefiting from more positive perceptions about their substantive and output legitimacy. In contrast, the various speculative financial applications of blockchain technology, through which many people have lost their savings, may not be perceived in the same way by external actors as they are by those who voluntarily take on the risks of such transactions.

As a general rule, the existence of both input and procedural legitimacy appears to be a common requirement for both endogenous and exogenous legitimacy. Yet, the values underlying endogenous legitimacy and exogenous legitimacy may differ.

Take, for example, the concept of *credible neutrality*. Buterin refers to the concept of credible neutrality as describing a situation where a large, heterogeneous group of people are convinced that a technical system does not discriminate for or against anyone. He provides the example of private property as a credibly neutral mechanism, because, in his view, it is a right enjoyed by all, and its existence “solves a lot of problems in society”.¹¹⁰ He then goes on to provide a set of principles for designing credibly neutral mechanisms, which are largely procedural in nature: e.g.,

¹⁰⁶ Ibid.

¹⁰⁷ See Johns (2023), Dimitropoulos (2022) and Walch (2019).

¹⁰⁸ See Reinsberg (2021).

¹⁰⁹ Reinsberg (2021), p. 304.

¹¹⁰ See Buterin (2020).

not writing specific people or outcomes into the mechanism, being open source, publicly verifying the execution of the software, keeping the mechanism simple, and not changing the technical rules of the mechanism very often—which he considers to be important features of blockchain-based systems.¹¹¹

In the political domain, neutrality, and the related concept of impartiality, are important public values necessary for the realization of the *rule of law* and the administration of justice. In that sense, the use of blockchain-based systems can help uphold rule of law principles.¹¹² However, other principles, such as equity—including the idea of equitable access and representation—are also important public values in many liberal democracies. In addition to civil rights legislation that prohibits discrimination on the grounds of race, nationality, skin color, religion, sex, and sexual orientation (e.g., US Civil Rights Act of 1964),¹¹³ there are many laws that seeks to empower traditionally disadvantaged and marginalized groups by positively discriminating in their favor (e.g., the Constitution of India allows the State to reserve civil service jobs for scheduled castes, tribes, and other “backward class of citizens”).¹¹⁴ The social and economic development of all citizens can thus be considered a public value, which allows for the meaningful use of equal opportunities—itsself another key liberal public value. Consequently, for a state that practices positive discrimination, the fair outcome from the use of a blockchain-based application may not necessarily be to treat everyone equally—as Buterin argues—but may rather require favoring certain groups. More broadly, we can conclude that political systems and Web3 ecosystems may have different criteria for determining the existence of input and procedural legitimacy, due to their differing value systems. Indeed, as Brummer has recently pointed out that in setting policies for financial regulation states may have conflicting policy goals (e.g., market integrity, financial innovation, and financial inclusion),¹¹⁵ all of which may not be able achievable at the same time. This clash or incompatibility in values can contribute to an unwillingness on the part of regulators and policy makers to recognize the regulatory equivalence of blockchain artifacts.

Moreover, there may be latent public values that have yet to be explicitly articulated. *Accountability*, for instance, is a public value that is central to public governance.¹¹⁶ However, the type of *accountability that is typically discussed presupposes the existence of a trusted authority that can be held to account*, such as a particular organ or entity of the state in the case of public governance. This understanding of accountability is difficult to transpose to a distributed network in which trusted authorities are eliminated, and in which holding network participants (e.g., core developers) individually account for harms arising from the operations of a

¹¹¹ Ibid.

¹¹² See De Vries and Kim (2011), De Filippi et al. (2022a)

¹¹³ Civil Rights Act of 1964 (Pub. L.88–352, 78 Stat. 241, enacted July 2, 1964) (United States).

¹¹⁴ Article 16(4), The Constitution of India, 1950.

¹¹⁵ See Brummer (2024).

¹¹⁶ See De Vries and Kim (2011).

technological system can lead to unjust outcomes.¹¹⁷ The vast majority of regulatory efforts targeted at the Web3 ecosystem have sought to identify and regulate the network participants that act as chokepoints (e.g., crypto-asset service providers in the EU's Markets in Crypto-Assets Regulation), or, when there are no apparent chokepoints, they have held some entities to be trusted authorities on weak grounds (e.g., the US OFAC sanction against the purported 'organization' behind the Tornado Cash software system).¹¹⁸ While this regulatory approach may be effective in deterring certain malevolent actors and transactions, it casts a wide net which could also inhibit socially and economically beneficial activities (e.g., open-source software development) due to the increased costs of compliance and risks of penalties.

As regulators typically have a difficult time mapping the myriad participants involved in a public, permissionless blockchain system, and identifying their particular roles in this system, they often treat them with suspicion. There is, therefore, a natural hesitation to recognize new ways of achieving public policy objectives through blockchain technology.

Take, for instance, the implementation of the DAO Model Law in the state of Utah in the United States. While Utah's DAO Act, for the most part, closely adheres to the text of the model legislation, there are certain key differences. One of these changes is the requirement that DAOs appoint a registered agent in the state of Utah to receive service of legal documentation and other communication from the State.¹¹⁹ Such a registered agent acts as an intermediary between the state and a legal entity, while also drawing local income from newly recognized DAOs. This choice reflects the need for state governments to identify an intermediary, and their desire for corporate entities to use local public and private services. In other words, the DAO Model Law provisions describing how blockchain technology could provide a technological equivalent of a registered agent requirement, for instance through direct service and notice to the DAO online, were not seen by legislators as being regulatorily equivalent to the public policies that animate the need to have registered agents. To uphold the public value of accountability, in the traditional conception of the term, a local physical connection between the DAO and the state had to be maintained.

In order to maximize the opportunities for regulatory equivalence to be recognized, one must not perceive the Web3 ecosystem to be in conflict with the political system, but rather in a competitive discovery process for facts that were previously unknown,¹²⁰ most notably concerning the public values that public actors seek to uphold but have not before been explicitly specified. As earlier literature critiquing the use of blockchain technologies in public administration has pointed out, there is a need for widely accessible forums of public debate to deliberate on major updates

¹¹⁷ See Haque et al. (2019).

¹¹⁸ See De Filippi and Mannan (2022).

¹¹⁹ Utah Code Ann. §48-5-101 – 406; § 48-5-106.

¹²⁰ See Hayek (2002).

and developments,¹²¹ as there is in other domains of public administration (e.g., building new highways). Such forums could help prevent the development of this technology (and the interpretation of laws relating to it) being guided by technical or domain experts but encourage the participation and input of laymen and policy-makers who may be otherwise hesitant to comment on technical issues.¹²²

The *discovery and explicit articulation of these values could strengthen co-regulatory efforts in the path towards achieving regulatory equivalence*. Indeed, in response to a clearer understanding of public values, it becomes possible to improve the technological design of blockchain systems in order to provide more advanced technological guarantees. Moreover, explicit public values can help shape the endogenous social norms within Web3 ecosystems, translating into voluntary (yet coordinated) efforts by web3 practitioners and industry players to address these public values through technological design.¹²³ For instance, in an effort to achieve both procedural and input legitimacy, Web3 ecosystems may develop Web3-native forms of accountability, which are more suitable to these new types of technological artifacts. Similarly, blockchain-based systems could be designed to incorporate rules that further human liberty and agency—values that are cherished in both traditional political regimes and Web3 systems.¹²⁴ Good faith efforts in this regard are likely to encourage more public authorities to see the merits of recognizing regulatory equivalence.

5 Conclusion

In this chapter, we first explained the concepts of *functional equivalence* and *regulatory equivalence*, distinguishing them from the notions of *technological neutrality* and *regulation by design* respectively. We then turned to the opportunities and challenges presented by blockchain artifacts as a means to achieve regulatory equivalence, noting that official recognition of this regulatory equivalence by lawmakers and public authorities is currently very challenging. The following section showed the centrality of public values to many major conceptions of legitimacy, drawing on various conceptions of legitimacy—as well as the interplay between endogenous and exogenous legitimacy—to show how the political system of public authorities and the Web3 ecosystem may have both convergent and divergent values. While substantive values are more likely to differ, both systems consider that procedural and input legitimacy should be attained. However, there may be differences in *how* such legitimacy should best be attained based on contrasting sets of values. The examples of ‘credible neutrality’ versus ‘positive discrimination’ were provided as

¹²¹ Hermstrüwer (2021), p. 112.

¹²² We thank Evrim Tan for this point.

¹²³ See De Filippi et al. (2022a).

¹²⁴ See De Filippi et al. (2022a).

potentially conflicting values between the Web3 ecosystem and liberal political systems. Similarly, different understandings of the concept of accountability across coercive centralized systems and non-coercive distributed systems might impinge upon the recognition that the Web3 ecosystem can also guarantee a degree of accountability. This clash of values can lead to inappropriate, even unjust, forms of regulation and foil attempts at having regulatory equivalence recognized.

Thus, for the full potential of blockchain technology and Web3 systems to be realized, there needs to be an ongoing dialogue between public authorities and the Web3 ecosystem. Indeed, the interplay between these two can be seen as a process for discovering and articulating implicit values that are not explicitly enshrined into the law, thereby contributing to enhancing mutual understanding and ideally bridging the gap between these divergent value systems.

Ultimately, recognizing the legitimacy of blockchain systems requires both sides to adapt: lawmakers must be open to new forms of accountability and transparency, while the Web3 community must strive to meet regulatory standards without compromising their foundational principles. Regulators providing more guidance on how Web3 communities can be made exogenously legitimate can enhance efforts from Web3 practitioners at finding ways to design blockchain-based systems in ways that would provide better regulatory equivalence. Only through this cooperative approach will it be possible to develop a regulatory framework that respects the innovative nature of blockchain technology while ensuring it aligns with broader societal values.

Acknowledgments We would like to thank the organizers and participants of the authors' workshop at the 'Public Governance and Emerging Technologies – Values, Trust, and Compliance by Design' conference, organized at Utrecht University, Utrecht, 11 January 2024, for their valuable comments. Usual disclaimers apply.

References

- Abreu PW, Aparicio M, Costa CJ (2018) Blockchain technology in the auditing environment. In: 2018 13th Iberian Conference on Information Systems and Technologies (CISTI). IEEE, pp 1–6
- Al-Jaroodi J, Mohamed N (2019) Blockchain in industries: a survey. *IEEE Access* 7:36500–36515
- Allena M (2022) Blockchain technology and regulatory compliance: towards a cooperative supervisory model. *Eur Rev Digit Adm Law* 2(2):37–44
- Almada M (2023) Regulation by design and the governance of technological futures. *Eur J Risk Regul* 14(4):697–709. <https://doi.org/10.1017/err.2023.37>
- Ametrano FM (2016) Bitcoin, blockchain, and distributed ledgers: between hype and reality. Available at SSRN 2832249
- Arredondo A (2018) Blockchain and certificate authority cryptography for an asynchronous online public notary system. Master's thesis, University of Texas at Austin
- Atzori M (2015). Blockchain technology and decentralized governance: Is the state still necessary? Available at SSRN 2709713. https://doi.org/10.22495/jgr_v6_i1_p5
- Beal J, Fisch B (2023) Derecho: privacy pools with proof-carrying disclosures. *Cryptology ePrint Archive*

- Bodó B, De Filippi P (2022) Trust in context: the impact of regulation on Blockchain and DeFi. Blockchain & Society Policy Research Lab Research Nodes, 2022(1)
- Bonsón E, Bednárová M (2019) Blockchain and its implications for accounting and auditing. *Meditari Account Res* 27(5):725–740
- Borgogno O (2022) Making decentralized autonomous organizations (DAOs) fit for legal life: mind the gap. Bank of Italy Occasional Paper (718)
- Brummer C (2024) Regulatory agencies and the Inclusion Trilemma. *Mod Law Rev* 87(1):1–23
- Buterin V (2020, January 3) Credible neutrality as a guiding principle. Nakamoto. <https://nakamoto.com/credible-neutrality/>
- Buterin V (2021, March 23) The most important scarce resource is legitimacy. Vitalik.ca. <https://vitalik.eth.limo/general/2021/03/23/legitimacy.html>
- Buterin V, Illum J, Nadler M, Schär F, Soleimani A (2023) Blockchain privacy and regulatory compliance: towards a practical equilibrium. Available at SSRN
- Cherednychenko O (2021) Two sides of the same coin: EU financial regulation and private law. *Eur Bus Org Law Rev* 22(1):147–172
- Choi C, De Filippi P, Dudley R, Elrifai SN, Fannizadeh F, Guillaume F, Leiter A, Mannan M, McMullen G, Riva S, Shimony O (2021) Model law for decentralized autonomous organizations (DAOs). Coalition of Automated Legal Applications
- Cirkovic M (2022). Cryptographic primitives for on-chain tumbler designs: improving the cost of autonomous cryptocurrency mixers. Bachelor's thesis, University of Bern
- Collomb A, De Filippi P, Sok K (2018) From IPOs to ICOs: the impact of Blockchain technology on financial regulation. Available at SSRN 3185347
- Collomb A, De Filippi P, Sok K (2019) Blockchain technology and financial regulation: a risk-based approach to the regulation of ICOs. *Eur J Risk Regul* 10(2):263–314. <https://doi.org/10.1017/err.2019.41>
- Cooper Z, Lodder AR (2023) The internet: what's law got to do with IT. In: Brožek B, Kanevskaia O, Pałka P (eds) *Research handbook on law & technology*. Edward Elgar Publishing, pp 45–58
- Cossar S, Reijers W, De Filippi P (2022) From liberal democracies to Blockchain systems: the instrumentalization of decentralization. In: Koskinen J, Kimppa KK, Heimo O, Naskali J, Ponkala S, Rantanen MM (eds) *Proceedings of the ETHICOMP 2022*. University of Turku, pp 315–328
- Dai J, Vasarhelyi MA (2017) Toward blockchain-based accounting and assurance. *J Inform Syst* 31(3):5–21
- Dalgetty E (2011) The role of a notary public in England and Wales. *Elder Law J* 3:324–327
- De Filippi P (2019) Blockchain technology and decentralized governance: the pitfalls of a trustless dream. *Decentralized Thriving: Governance and Community on the Web*, 3.0, hal-02445179
- De Filippi P (2021) Blockchain technology as an instrument for global governance. *Digital, Governance and Sovereignty Chair*, 1–16
- De Filippi P, Hassan S (2018) Blockchain technology as a regulatory technology: from code is law to law is code. arXiv preprint arXiv:1801.02507
- De Filippi P, Mannan M (2022, December 6). Tornado Cash: the end of Blockchain neutrality? Zora Zine. <https://zine.zora.co/tornado-cash-primavera-de-filippi-morshed-mannan>
- De Filippi P, Wright A (2018) Blockchain and the law: the rule of code. Harvard University Press
- De Filippi P, Mannan M, Reijers W (2020) Blockchain as a confidence machine: the problem of trust & challenges of governance. *Technol Soc* 62:101284
- De Filippi P, Mannan M, Reijers W (2022a) Blockchain technology and the rule of code: regulation via governance. *George Wash Law Review* 92. forthcoming
- De Filippi P, Mannan M, Reijers W (2022b) The alegality of blockchain technology. *Policy Soc* 41(3):358–372
- De Filippi P, Mannan M, Henderson J, Merk T, Cossar S, Nabben KA (2022c) Report on blockchain technology & legitimacy. RSC Research Report. <https://hdl.handle.net/1814/75167>
- de Vries MS, Kim PS (2011) Introduction to value and virtue in public administration. In: de Vries MS, Kim PS (eds) *Value and virtue in public administration*. Palgrave Macmillan

- Dias Menezes L, Vieira de Araújo L, Nishijima M (2023) Blockchain and smart contract architecture for notaries services under civil law: a Brazilian experience. *Int J Inform Secur* 22:869–880
- Dimitropoulos G (2022) The use of blockchain by international organizations: effectiveness and legitimacy. *Policy Soc* 41(3):328–342
- Egan M (2002) Setting standards: strategic advantages in international trade. *Bus Strategy Rev* 13(1):51–64
- Ferrari V (2020) The regulation of crypto-assets in the EU: investment and payment tokens under the radar. *Maastricht J Eur Comp Law* 27(3):325–342
- Finocchiaro G, Castellani L (2023) The UNCITRAL model law on electronic transferable records: introduction and overview. *Eur Rev Priv Law* 31(5):955–978
- Gertz N, Ozkaramanli D (2024) Beauvoir versus behavior change: introducing existential ethics to the politics of design. *Design Issues* 40(2):56–67
- Gonzalez R (2023) The impact of DAOs on corporate law: an analysis of DAO frameworks and potential legal implications. Featured Student Work, University of San Francisco
- Guillaume F (2023) Decentralized autonomous organizations (DAOs) before state courts: how can private international law keep up with global digital entities? In: Perestrelo de Oliveira M, Garcia A (eds) *DAO regulation: principles and perspectives for the future*
- Haque RS, Silva-Herzog RS, Plummer BA, Rosario NM (2019) Blockchain development and fiduciary duty. *Stanf J Blockchain Law Policy* 2(2):139–188
- Harvey DJ (2017) *Collisions in the digital paradigm: law and rule-making in the internet age*. Hart Publishing
- Hassan S, De Filippi P (2021) Decentralized autonomous organization. *Inter Policy Rev* 10(2):1–10
- Hayek F.A (2002) Competition as a discovery procedure in the quarterly journal of Austrian economics (5)3: 9–23. Translated by marcellus s. snow from *Der Wettbewerb als Entdeckungsverfahren*. (1968 lecture)
- Henning J (2018) The Howey test: are crypto-assets investment contracts? *Univ Miami Bus Law Rev* 27:51–74
- Herian R (2018) *Regulating Blockchain: critical perspectives in law and technology*. Routledge
- Hermstrüwer Y (2021) Blockchain and public administration. In: Pollicino O, De Gregorio G (eds) *Blockchain and public law: global challenges in the era of decentralisation*. Edward Elgar Publishing, pp 104–121
- Johns F (2023) #Help: digital humanitarianism and the remaking of international order. Oxford University Press
- Kerr DS, Loveland KA, Smith KT, Smith LM (2023) Cryptocurrency risks, fraud cases, and financial performance. *Risks* 11(3):51
- Koos S (2023) The digitization of notarial tasks: a comparative overview and outlook of ‘Cyber Notary’ In Indonesia and Germany. *Indonesian J Socio-Legal Stud* 2(2):1. <https://doi.org/10.54828/ijsls.2023v2n2.1>
- Lemieux VL, Hofman D, Batista D, Joo A (2019) *Blockchain technology & recordkeeping*. ARMA International Educational Foundation
- Levi M (2019) Trustworthy government and legitimating beliefs. *Nomos* 61:362–384
- Magnuson W (2018) Regulating fintech. *Vanderbilt Law Rev* 71(4):1167–1215
- Mahmoud AIA (2022) Exploring the potential of DAOs: a comprehensive study of decentralized autonomous organizations. Master’s thesis, Politecnico Milano
- Mannan M (2023) The promise and perils of corporate governance-by-design in Blockchain-based collectives: the case of dOrg. In: Manley J, Webster A, Kuznetsova O (eds) *Co-operation and co-operatives in 21st-century Europe*. Bristol University Press, pp 78–99
- Mason S (2018) Documents signed or executed with electronic signatures in English law. *Comput Law Secur Rev* 34(4):933–945. <https://doi.org/10.1016/j.clsr.2018.05.023>
- Mazgaj MT (2021) Blockchain-based registration and transfer of shares: consequences in the area of the law of obligations. In: Slakoper Z, Tot I (eds) *Digital technologies and the law of obligations*. Routledge, pp 77–94

- Meijer D, Ubacht J (2018) The governance of blockchain systems from an institutional perspective: a matter of trust or control? In: *Proceedings of the 19th Annual International Conference on Digital Government Research: Governance in the Data Age*. Association for Computing Machinery, pp 1–9
- Micheler E, Whaley A (2020) Regulatory technology: replacing law with computer code. *Eur Bus Org Law Rev* 21:349–377
- Miscione G, Kavanagh D (2015). Bitcoin and the Blockchain: a coup d'état through digital heterotopia?. *Humanistic Management Network Research Paper Series*, (23/15)
- Morrison R, Mazey NC, Wingreen SC (2020) The DAO controversy: the case for a new species of corporate governance? *Front Blockchain* 3:25
- Nastri M (2021) Blockchain, smart contracts and new certainties: what future for notaries? In: Cappiello B, Carullo G (eds) *Blockchain, law and governance*. Springer International Publishing, pp 221–228
- Parsons T (1960) *Structure and process in modern societies*. Free Press
- Pereira J, Tavalaei MM, Ozalp H (2019) Blockchain-based platforms: Decentralized infrastructures and its boundary conditions. *Technol Forecast Soc Change* 146(C):94–102
- Putz B, Menges F, Pernul G (2019) A secure and auditable logging infrastructure based on a permissioned blockchain. *Comput Secur* 87(7-8):101602
- Raskin M (2016) The law and legality of smart contracts. *Georgetown Law Technol Rev* 2017(1):305–341
- Reed C (2010) Online and offline equivalence: aspiration and achievement. *Int J Law Inform Technol* 18(3):248–273
- Reinsberg B (2021) Fully-automated liberalism? Blockchain technology and international cooperation in an anarchic world. *Int Theory* 13(2):287–313
- Renwick R, Gleasure R (2021) Those who control the code control the rules: how different perspectives of privacy are being written into the code of blockchain systems. *J Inform Technol* 36(1):16–38
- Reyes CL (2019) If Rockefeller were a coder. *George Wash Law Rev* 87:373–401
- Rogers B (2023) *Data and democracy at work: advanced information technologies, labor law, and the new working class*. MIT Press
- Rosati P, Lynn TG, Fox G (2021) Blockchain: a technology in search of legitimacy. In: Patnaik S, Wang TS, Shen T, Panigrahi SK (eds) *Blockchain technology and innovations in business processes*. Springer, pp 17–32
- Rozenshtein AZ (2018) Surveillance intermediaries. *Stanf Law Rev* 70(1):99–189
- Savin A (2019) Rule making in the digital economy: overcoming functional equivalence as a regulatory principle in the EU. *J Inter Law* 22(8):3–15
- Shin D, Hwang Y (2020) The effects of security and traceability of blockchain on digital affordance. *Online Inform Rev* 44(4):913–932
- Silva Ramalho, D., & Igreja Matos, N. (2021) What we do in the (digital) shadows: anti-money laundering regulation and a bitcoin-mixing criminal problem. In *ERA Forum* 22, 3 (pp. 487–506). Springer Berlin Heidelberg.
- Sims A (2019) Blockchain and decentralised autonomous organisations (DAOs): the evolution of companies? *New Zealand Univ Law Rev* 28:423–458
- Spouse T, Green B (2019) How distributed ledger technology could solve regulatory problems. *J Financ Compl* 3(1):60–66
- Sulkowski AJ (2019) The Tao of DAO: hardcoding business ethics on blockchain. *Bus Financ Law Rev* 3(2):146–166
- Sullivan C, Burger E (2017) E-residency and blockchain. *Comput Law Secur Rev* 33(4):470–481
- Taddeo M, Floridi L (2016) The debate on the moral responsibilities of online service providers. *Sci Eng Ethics* 22(6):1575–1603
- Temte MN (2019) Blockchain challenges traditional contract law: just how smart are smart contracts. *Wyoming Law Rev* 19(1):87–111

- Thakur A (2023) Blockchain and cryptocurrency frauds: emerging concern. In: Gupta CM (ed) *Financial crimes: a guide to financial exploitation in a digital age*. Springer, pp 131–145
- Tse N (2020) Decentralised autonomous organisations and the corporate form. *Victoria Univ Wellington Law Rev* 51(2):313–356
- Ulloa R, Gallegos P (2021) Design of a Blockchain architecture and use of smart contracts to improve processes in notary office. In: Botto-Tobar M, Montes León S, Torres-Carrión P, Zambrano Vizuete M, Durakovic B (eds) *3rd international conference on applied technologies*. Springer, pp 469–483
- UNCITRAL (1996) *UNCITRAL model law for electronic commerce and guide to enactment*. United Nations
- Veerpalu A (2019) Functional equivalence: an exploration through shortcomings to solutions. *Baltic J Law Polit* 12(2):134–162
- Vlcek W (2008) A Leviathan rejuvenated: surveillance, money laundering, and the war on terror. *Int J Polit Cult Soc* 20(1–4):21–40
- Walch A (2019) In code(rs) we trust: software developers as fiduciaries in public Blockchains. In: Hacker P, Lianos I, Dimitropoulos G, Eich S (eds) *Regulating Blockchain: techno-social and legal challenges*. Oxford University Press, pp 58–59
- Wang Z, Xiong X, Knottenbelt WJ (2023) Blockchain transaction censorship: (In)secure and (in) efficient?. *Cryptology ePrint Archive*
- Weber M (1964) *The theory of social and economic organization*. Free Press
- Werbach K (2018) Trust, but verify: why the Blockchain needs the law. *Berkeley Technol Law J* 33(2):487–550
- Werbach K, Cornell N (2017) Contracts Ex Machina. *Duke Law J* 67(2):313–384
- Wood G [POW Media] (2015, May 14). Alegality: systems that can't care. CoinScrum and proof of work media: tools for the future: London 2014 [Video]. YouTube. <https://www.youtube.com/watch?v=Zh9BxYTSrGU>
- Yeoh P (2017) Regulatory issues in blockchain technology. *J Financ Regul Compl* 25(2):196–208
- Yeung K (2019) Regulation by Blockchain: the emerging battle for supremacy between the code of law and code as law. *Mod Law Rev* 82(2):207–239
- Zhou M, Shi E, Fanti G (2023) Proof of compliance for anonymous, unlinkable messages. *Cryptology ePrint Archive*
- Zwitter A, Hazenberg J (2020) Decentralized network governance: blockchain technology and the future of regulation. *Front Blockchain* 3:12

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.



The Need for Responsible Use of AI by Public Administration: Algorithmic Impact Assessments (AIAs) as Instruments for Accountability and Social Control



Julia Iunes Monteiro

Abstract AI is increasingly being implemented in the decision-making process of public administration, from the provision of social benefits to public security policies, generating wide impacts on citizen's fundamental rights. It is expected that governments' use of AI will increase the efficiency of public management. Nonetheless, its indiscriminate use is undermining the legality, transparency, and justification of administrative acts. An increasingly necessary question in this new realm is how to guarantee accountability for algorithmic policies implemented by the state. What duties should the public administration assume when implementing AI technologies that impact fundamental rights? Although regulation and practice are still early, algorithmic impact assessments (AIAs) are gaining prominence among academics and policymakers as a key accountability mechanism that should be applied before introducing an algorithmic system in a given context, to prevent its risks, and evaluate its societal implications. Given the importance that AIAs can have for the preservation of trust in government, the rule of law, and fundamental rights, the general aim of this contribution is to explore the fundamentals of AIA, indicating the challenges, stages, and requirements for it to become an effective mechanism for algorithmic accountability by the public administration.

J. Iunes Monteiro (✉)

Pontifical Catholic University of Rio de Janeiro (PUC-Rio), Rio de Janeiro, Brazil

© The Author(s) 2025

J. Goossens et al. (eds.), *Public Governance and Emerging Technologies*,
https://doi.org/10.1007/978-3-031-84748-6_9

179

1 Introduction

The public administration¹ of several countries is increasingly applying artificial intelligence (AI) to assist various activities, such as public services provision, decision-making processes, and law enforcement.² Although the use of AI by state actors is still in its inception compared with the private sector, it has been provoking wide-ranging impacts on fundamental rights.

Automated systems are influencing access to social benefits without proper accuracy and due process guarantees, as was the case with SyRI (“*system risk indication*”) deployed by the Dutch government and systems for social security provision implemented by the federal government of Brazil. Moreover, algorithms are also used for criminal prosecution activities, through facial recognition and predictive policing technologies, introduced in several cities worldwide.

It is expected that governments’ use of AI will increase the efficiency of public management. Nonetheless, its indiscriminate use is causing damage to society as a whole and to the principles of legality, equality, and public justification that should underpin state acts.³ There is overwhelming evidence of the discriminatory potential of algorithms, especially for groups already marginalized in society.⁴ Furthermore, its lack of transparency and explainability often undermines citizens’ ability to understand, challenge, and scrutinize public decisions that affect their lives.⁵

Despite the systemic impact that algorithms can have on society, they are being implemented by the public sector from the Global North and South without a proper assessment of their risks. Due to the lack of specific accountability mechanisms, states rarely give a public justification for the adoption of these technologies, explain how they work, and take responsibility for eventual damage they may cause.⁶

An increasingly necessary question in this new realm is how to guarantee accountability for algorithmic policies implemented by the state. What duties should the public administration assume when implementing algorithmic technologies that impact fundamental rights? Recent research shows that regulatory frameworks for public accountability are still in their infancy, and governments generally struggle to implement AI responsibly.⁷

Although the regulatory efforts are still early, algorithmic impact assessments (AIAs) are gaining prominence among academics and policymakers as one of the

¹The term public administration is used in reference to the typical administrative function performed by the Executive Branch, in managing the public interest and applying the law. Therefore, the object of our analysis is AI systems deployed in the decision-making process of public administration to assist in law enforcement activities and in the provision of public services to society.

²See Fernández (2022).

³See Smuha (2021) and Nemitz (2018).

⁴See Eubanks (2018); O’Neil (2016); Buolamwini and Gebru (2018).

⁵See Danaher (2016).

⁶See Brandusescu and Reia (2022) and Reismann et al. (2018).

⁷See Ada Lovelace et al. (2021) and Moss et al. (2021).

main accountability mechanisms that should be applied *ex ante*, before introducing an algorithmic system in a given context. AIA originates in impact assessment processes already developed in other fields, such as Data Protection Impact Assessment (DPIA) and Environmental Impact Assessment.

Inspired by these existing models, AIA aims to enable a prior and participatory study of algorithmic risks, costs, benefits, and societal implications. It is recommended that AIA's processes include meaningful civil society participation. Moreover, it is hoped that, when adequately institutionalized, AIAs may inform public agencies' decisions on whether to adopt the technological system and how to mitigate its potential harms.

Despite the existence of some regulations and research on AIA, "there is still no universally accepted standard on its execution methodology", especially when implemented by the public administration.⁸

Although there are already some ongoing regulatory and practical initiatives worldwide, AIAs are still a relatively new and poorly institutionalized mechanism, especially regarding their application by the public administration. AIAs are being developed inspired by impact assessment models already present in other areas, such as environmental and data protection impact assessments. However, there is still uncertainty about in which cases and by which processes AIAs should be conducted.

Considering the European context, assessments to evaluate the impact of algorithms on fundamental rights were established at a regional level only in 2024, through the AI Act. At the moment, little is known about how the Fundamental Rights Impact Assessment (FRIA) provided by the AI Act will be implemented in practice, as it will be mandatory only from August 2026. Furthermore, the practical implementation of FRIAs will also depend on how the regulatory authorities of each Member State of the EU will supervise and complement the current provisions of the AI Act in the future.

On the other hand, if we consider the Latin American context, a proper institutionalization of AIAs seems to be even further away. According to a recent report by Access Now,⁹ none of the Latin American countries have national regulations on artificial intelligence. This raises concerns about the proper institutionalization and implementation of AIAs in these countries.

In Brazil, for instance, the importance of AIAs was only acknowledged by regulatory discussions from 2023 onward. At that time, an amendment to the ongoing AI draft bill was proposed to make impact assessments a mandatory procedure for high-risk AI systems.¹⁰

Given the recent regulation of AIAs in Europe and the emerging regulatory debate in Latin America, it seems conceivable to deduce that this mechanism is in

⁸Lemos et al. (2023), p. 26.

⁹See Access Now (2024).

¹⁰See Lemos et al. (2023).

its early stages of institutionalization, especially considering the Global South and the “majority world”.¹¹

In addition, the particularities of AIA when implemented by public bodies are still unclear, and there are relatively few studies on this topic, as verified in our review of publications in English and Portuguese. As stated by Mantelero (2024), the methodological debate around AIAs is limited and there is a “lack of experience with HRIA [Human Rights Impact Assessments] in relation to AI”.¹²

Given the importance that AIAs can have for the preservation of trust in government, the rule of law, and fundamental rights, the general aim of this contribution is to explore the fundamentals of AIA, indicating the requirements and challenges for it to become an effective mechanism for algorithmic accountability by the public administration.

To accomplish the aforementioned goals, this contribution is divided into four sections. The first section provides context on the current state of AI implementation by governments, highlighting the absence of accountability mechanisms provided by the public administration to enable proper democratic and social control of AI-based policies and decisions. The second section presents AIAs as a crucial accountability measure to be implemented, outlining the main objectives an AIA should pursue when deployed by the public administration. The third section explores the essential criteria for conducting an AIA, its relationship with impact assessments applied in other fields (environmental, data protection, and public policy), and current efforts to regulate this instrument.

The fourth and final section concludes by stating some essential challenges for AIA to become an effective mechanism for public accountability. A crucial element that will deserve more attention from the literature and public policymakers is the introduction of processes for civil society to participate in impact assessments.

The methodology used is a review of legal and sociotechnical literature on AIA, with a focus on AIAs designed specifically for the public sector.¹³ The literature review aimed to comprehend the primary academic recommendations for regulating and institutionalizing AIAs in public administration processes.

In addition, a document analysis of national and foreign regulations on the subject was also carried out, with a specific focus on the AI Act, recently approved in the European Union, and the draft bill for a Legal Framework for AI Regulation, which is being debated in Brazil. The AI Act was chosen for being the first and more

¹¹ Majority world is a term coined by Shahidul Alam, a writer and activist from Bangladesh that proposes a shift of standpoint from Global North perspectives to perspectives that consider the needs, knowledge, innovation, and labor of the majority of the human population. See Amrute et al. (2022).

¹² Mantelero (2024), p. 3.

¹³ The literature on Algorithmic Impact Assessments refers to the “public sector” as a way of distinguishing impact assessments conducted by public and private actors. Although the term “public sector” encompasses the three branches of government, this work is concerned with the study of AIAs in the Executive branch, namely, the public administration.

comprehensive effort to regulate AI among global economies and the first regulatory framework for AI in the European Union.¹⁴

Conversely, Brazil has been advancing the regulatory discussion on AI since 2021. The primary proposal being deliberated in the Brazilian National Congress is rather inspired by the AI Act. Furthermore, Brazil is recognized as one of the Latin American countries with the most advanced regulatory discussion on AI,¹⁵ and there are expectations that its national regulation on AI will be soon approved.

Therefore, the purpose of analyzing these regulatory initiatives is to exemplify whether and how academic recommendations on AIA criteria and stages are being integrated into current efforts to regulate AIAs in the public sector.

2 The Use of AI in Public Administration and the Need for Algorithmic Accountability Measures

Various countries' public administrations currently utilize algorithmic systems to assist their decision-making processes. These systems range from simpler rule-based algorithms that automate the application of traffic fines¹⁶ to more sophisticated machine-learning systems capable of learning from the environment and making predictions. The latter are generally referred to as artificial intelligence, precisely because they are capable of performing tasks normally associated with human intelligence.¹⁷

There is no fixed rule as to which types of algorithms should be classified as "artificial intelligence". Nonetheless, AI can be generally understood as an umbrella term for various computer models designed by humans to perform complex data analysis, risk prediction, and rule-making activities. These systems can process vast amounts of data, categorize information, and apply mathematical models to decide on the actions to achieve the objectives set.¹⁸

The ability of AI to provide objective answers to complex problems is celebrated. Nonetheless, this does not automatically make them neutral, fair, or impartial. Human biases and political views are embedded in algorithms' design, as humans define their goals, pre-classify the training data, and adjust their limits and parameters.¹⁹ Due to the social implications of these systems, there is currently a

¹⁴The AI Act was proposed by the European Commission in April 2021 to create the first regulatory framework for AI in the European Union. This initiative aims to unify AI regulation across the EU's 27 member states. See Mökander et al. (2021).

¹⁵See Oxford Insights (2023).

¹⁶See Zouridis et al. (2019).

¹⁷See Smuha (2022).

¹⁸European Commission (2019), p. 36.

¹⁹See Burrell (2016).

worldwide movement to regulate them in such a way as to protect human rights, democracy, and the rule of law.

Given the potential of AI to analyze data and inform the most varied types of decisions, there is an expectation that these systems can help governments provide more accurate and efficient public services and policies.²⁰ On the other hand, algorithmic systems often amplify discriminatory biases and present problems of opacity. Operating as “black boxes”²¹ opaque to public scrutiny, algorithmic decisions pose even deeper obstacles to citizens’ right to understand, monitor, and challenge state decisions that affect their lives.

One example is facial recognition technology,²² which is used in several cities worldwide to guide the work of the police in identifying and prosecuting suspects. Studies have shown that facial recognition systems have lower accuracy in identifying black people. This leads to higher rates of false positives, which in turn exacerbates police violence and discrimination against these communities.²³

Problems of opacity and algorithmic discrimination have also occurred with predictive policing systems, such as those implemented in the USA and Chile²⁴ and automated systems used by the public administration to provide access to social benefits. One example is SyRI (“*system risk indication*”) implemented by the Dutch government to identify potential fraudulent beneficiaries, which was heavily criticized and consequently banned by the Hague District Court for violating Article 8 of the European Convention on Human Rights (ECHR), which protects the right to respect for private and family life, home and correspondence. According to the court, the Netherlands did not strike the right balance between the benefits such technology brings and the infringement of the right to a private life through the use of emerging technologies.²⁵

Research shows that not only was the algorithm unsuccessful in detecting new fraud cases, but it also contributed to the expansion of class biases and xenophobia, as it was primarily used in low-income neighborhoods, feeding back into the system’s chances of identifying fraud risks exclusively in the poorest areas, mostly inhabited by immigrants.²⁶

The examples listed above demonstrate how AI has been implemented by the public sector in the Global North and South without a proper prior assessment of its

²⁰ See Worldbank (2021).

²¹ See Pasquale (2015).

²² Facial recognition systems process facial biometric data collected from images and videos. The algorithms can identify biological patterns in the images based on mathematical calculations, such as the distance between the eyes and the shape of the nose. Once these patterns have been identified, the systems compare the captured images with existing images in a previous database, thus helping to identify wanted individuals.

²³ See Buolamwini and Gebru (2018), Rodrigues (2019), Costa and Kremer (2022).

²⁴ See Buschmann (2021).

²⁵ See van Bekkum and Borgesius (2021).

²⁶ See Algorithm Watch (2020), Amnesty International (2021) and Newman and Mintrom (2023).

risks. As widely reported by the smart city literature,²⁷ the deployment of technology by governments is often driven by a techno-solutionist mentality and by the transference of public functions to private companies, which generally develop these systems in a way that disregards the necessities and particularities of local contexts.

Studies conducted by research institutes and civil society organizations also reveal that the use of AI in the public sector is generally related to “ill-designed public policies”.²⁸ Either because the formulators of these policies disregard the potential impacts of automated systems on human rights or because they overestimate the system’s ability to “solve” complex social problems, such as the problem of public security, for example, which unequivocally demands political-institutional transformations that go far beyond the reach of an algorithm.²⁹

In addition to a techno-solutionist mindset that often underpins algorithmic policies, a lack of public information, justification, and deliberation about these policies is worrisome. As algorithms are, in many cases, acting as the law itself,³⁰ their design and implementation should follow parameters of justification and political deliberation that underpin state acts.³¹ Before applying technology to any social problem, governments should consider *how*, *when*, and *why to* implement algorithms in the first place. It encompasses the consideration of which social issues should be prioritized by the government, what would be the most appropriate approach to tackle the issue at hand, which groups of people would benefit from this approach, and whether the use of technology is in fact needed to achieve the desired ends.³²

Despite the need for prior public debate and disclosure of information about algorithmic policies, a report by the AI Now Institute (2018) reveals that “currently, few public agencies are explicitly obliged to disclose anything about the systems they have or plan to use”,³³ leaving society to access the data that journalists, research institutes and human rights organizations can expose, based on isolated investigative efforts, requests for access to information and independent audits.

Not only is the lack of information on these policies troubling, but so is the absence of prior evaluations to measure their risks and attest to their necessity and adequacy.³⁴ Surveillance and criminal prosecution systems, such as those mentioned

²⁷ See Morozov and Bria (2018), Jewell (2018) and Lynch (2020).

²⁸ Brandusescu and Reia (2022).

²⁹ As Krasmann (2020) points out, algorithms employ a “superficial logic”, reducing our view of reality to what the data can show and the algorithms can calculate.

³⁰ The idea that algorithms have “normative power” over society goes back to the seminal work of Lawrence Lessig (1999), “Code: And Other Laws of Cyberspace” which, back in the 1990s, pointed to the ability of algorithmic systems to actually ‘govern’ human behavior. See Lessig (1999).

³¹ See Smuha (2022) and Fernandez (2022).

³² See Brandusescu and Reia (2022).

³³ Reisman et al. (2018), p. 4.

³⁴ This problem is constantly reported in research and external audits carried out by research centers and civil society organizations. A case study on the Urban Crime Predictive System imple-

above, have been the target of intense protest by social movements due to their systematic violations of human rights,³⁵ which has led to successive bans³⁶ and prohibitions of these technologies in various parts of the world.

In many cases, the negative consequences resulting from the use of automated systems are only discovered after they have already caused significant harm to individuals and communities, as well as to the legitimacy of democratic processes and the public budget. Civil society groups argue that funds allocated for such initiatives could be used for other policies that would better benefit the population.³⁷

The current scenario of AI implementation shows that governments urgently need to move away from a techno-solutionist approach towards a responsible and precautionary³⁸ approach when adopting algorithmic policies. A responsible approach requires that public agencies aiming to implement systems with likely impacts on fundamental rights develop *accountability* parameters, such as algorithmic impact assessments, to justify the adoption of these technologies, explain how they work, attest to their benefits, and take responsibility for any possible damage. But how can accountability be applied to algorithmic policies? And why is AIA an important tool in this regard?

mented by the Chilean federal government revealed that none of those responsible for developing the system considered its ethical implications and its impact on human rights. See Buschmann (2021) and Brandusescu and Reia (2022).

³⁵In Brazil, for example, the organization “Tire Meu Rosto da Sua Mira” is calling for a total ban on facial recognition technologies for public security purposes. In the US, organizations such as the Movement for Black Lives also advocate an end to surveillance and predictive policing systems and the “detour of public funds used for surveillance to meet community needs”. More information, respectively, at: <https://tiremeurostodasuaamira.org.br/> and <https://m4bl.org/policy-platforms/end-surveillance/>.

³⁶Faced with public pressure on the adoption of population surveillance technologies, several local governments have enacted legislation banning the use of AI and facial recognition technologies for public safety purposes. This is the case, for example, in San Francisco (USA) and 16 other US municipalities. See Ghaffary (2019) and Sheard and Schwartz (2022).

³⁷In the US, organizations such as the Movement for Black Lives also advocate an end to surveillance and predictive policing systems and the “detour of public funds used for surveillance to meet community needs”. See <https://m4bl.org/policy-platforms/end-surveillance>.

³⁸I indicate the need for a precautionary approach in reference to the precautionary principle, which, as indicated by Bioni and Luciano (2020), demands “public deliberation on whether or not to adopt AI in view of the definition of the type of risks that underlie them”.

2.1 *Algorithmic Accountability in the Context of Public Administration*

To use Mark Bovens' widely quoted definition, accountability consists of "the relationship between an actor and a forum, in which the actor has an obligation to explain and justify his conduct, the forum can ask questions and make judgments, and the actor can face consequences".³⁹

In other words, accountability is not limited to explanations of the *actor* about its conduct but necessarily requires the existence of an *external forum* that can evaluate and express its opinion on the adequacy of the information and justifications provided. In addition, accountability also involves the possibility of making the actor responsible for the consequences of their action. This can happen in a variety of ways, whether through denunciations, public condemnation of their behavior, redress mechanisms, and the possible application of legal sanctions.

Although accountability systems can vary, the important thing is that they involve the following three elements: (i) justification of the conduct, (ii) judgment by an external forum, and (iii) the attribution of consequences.

In the context of public administration, accountability is closely related to democratic and social control. It involves *horizontal accountability*, carried out by other branches of government that supervise public administration, as well as *social accountability*, implemented by citizens and civil society organizations.⁴⁰ By providing information and justifications for its conduct, the government enables parliament and society to monitor whether its actions align with its stated objectives and the public interest.

Concerning the use of AI by the public sector, *algorithmic accountability*⁴¹ consists of the duty of public agencies to justify the adoption of algorithmic systems, explain their functioning and decisions, and take responsibility for their potential effects on individuals and society. It is recommended that accountability mechanisms be placed throughout the "AI life cycle", from their development and/or acquisition to their implementation and evaluation.⁴² Accountability instruments can be diverse, ranging from mechanisms for public transparency to periodic audits to processes for explaining and challenging AI-based decisions.

In the last few years, AIA has stood out as an important accountability mechanism being gradually applied by governments in different jurisdictions. Research

³⁹ Bovens (2007), p. 450.

⁴⁰ See O'Donnell (2004).

⁴¹ The concept of algorithmic accountability is discussed in a systematic review by Wieringa (2020), who states that: "Algorithmic accountability concerns a networked account for a socio-technical algorithmic system, following the various stages of the system's lifecycle. In this accountability relationship, multiple actors (e.g. decision makers, developers, users) have the obligation to explain and justify their use, design, and/or decisions of/concerning the system and the subsequent effects of that conduct". See Wieringa (2020), p. 10.

⁴² See Wieringa (2020).

conducted by the Ada Lovelace Institute in partnership with the AI Now Institute and the Open Government Partnership (2021) mapped what they called the “first wave” of algorithmic accountability policies in the public sector. Their research concludes that AIA is one of the main accountability measures being adopted by governments, based on an analysis of more than 40 local and national cases in more than 20 different jurisdictions, including Chile, Canada, the United Kingdom, the Netherlands, France, New Zealand, Amsterdam, New York, and Seattle.

Conducting AIA can inform public bodies on whether to adopt the system, based on the risks identified and the potential institutional and social impacts that the system will generate in the future.

Due to its potential to prevent future risks generated by AI, algorithmic impact assessments have been considered a central element of a risk-based approach to AI regulation. Consequently, they are being increasingly studied by academics, recommended by policymakers, and included both in existing regulatory initiatives, such as the AI Act recently approved on 13 March in the European Union, and in regulatory projects that are still underway, such as the Legal Framework for Artificial Intelligence, which is being debated in Brazil.⁴³

When applied by the public sector, it is expected that algorithmic impact assessment can contribute to the formulation of better-informed algorithmic policies, to the protection of fundamental rights, and the social control of public acts. Its foundations, objectives, and application criteria will be addressed in the following topics.

3 The Goals of Algorithmic Impact Assessments (AIA) in the Public Administration

The report “Assembling Accountability: Algorithmic Impact Assessment for the Public Interest”, produced by the Data & Society Research Institute (2021) clarifies that algorithmic impact assessments seek to answer three main questions: “what a system does; who can do something about what the system does; and who should make decisions about what the system can do”.⁴⁴ By answering these questions, impact assessments “make it possible to simultaneously document an undertaking, assess the impacts it may cause, and assign responsibility for those impacts”.⁴⁵

Although AIA is still in the early stages of being widely institutionalized and implemented, it is being considered a fundamental process for any organization, whether public or private, that intends to develop or use algorithmic systems

⁴³The Legal Framework for the Regulation of Artificial Intelligence in Brazil has been debated by the Brazilian Legislative Branch since 2021, and is currently awaiting a vote in the National Congress. See Agência Senado (2024). For a general analysis on how AIA is regulated by the Brazilian proposal, see Mulholland et al. (2023).

⁴⁴Moss et al. (2021), p. 5.

⁴⁵Moss et al. (2021), p. 10.

affecting fundamental rights.⁴⁶ However, when AI is applied by the public sector, AIAs become even more essential, given the constitutional duty of states to be accountable to the population and enable democratic and social control of public acts.

Therefore, AIA in the public sector aims to be a mechanism for government accountability regarding the implementation of algorithmic policies. An impact assessment therefore consists of a preliminary study through which public bodies can assess the risks, costs, benefits, and socioeconomic impacts of implementing the AI system in a given institutional and/or social context.

The information provided by government bodies as part of this preliminary study must not only enable technical scrutiny of the algorithms but also an assessment of the political and social suitability of their implementation.

Based on the literature review conducted, it is possible to state that algorithmic impact assessment carried out by the public sector should fulfill five main purposes: (i) information, (ii) risk assessment (preventive and regulatory function), (iii) weighting of cost and benefits, (iv) public participation and (v) accountability.

3.1 Information

An impact assessment has an *informative* purpose because its initial objective is to gather information about the algorithmic system, its operation, and potential consequences. This information is necessary for two main reasons. First, it allows democratic and social oversight of the algorithmic policies put in place by the public administration. Secondly, it ensures that the public administration understands the potential risks and impacts of the systems it either develops or acquires from private technology companies. Thus, the informative purpose of impact assessments has two dimensions: to inform the government agencies implementing the system, as well as to inform society and the other branches of government responsible for external control of public administration.

As indicated by Reisman and colleagues, algorithmic impact assessments conducted by public bodies should be conceived as “practical tools to inform the political debate on the use of these systems and provide information to communities that can help determine whether these systems are appropriate”.⁴⁷ Therefore, the informative purpose of impact assessments is grounded on the *duty of transparency* of state acts and on the right of the people subjected to AI systems to *access information* about their operation and about how it can influence their fundamental rights.

The *information* generated during the impact analysis can be used to fulfill the other functions of *risk assessment*, *weighting*, *public participation*, and *accountability*.

⁴⁶ See Lemos et al. (2023), Moss et al. (2021) and Selbst (2021).

⁴⁷ Reisman et al. (2018), p. 6.

3.2 Risk Assessment (Preventive and Regulatory Function)

A second purpose of AIA is to promote an *assessment of the system's risks*. This risk assessment may have a preventive and regulatory function. AIA has a preventive function because its aim is to identify in advance any damage that could be generated by the system so that it can be prevented or mitigated before it is developed or deployed. As stated by Selbst, the first goal of an AIA is to “get the people who build systems to think methodically about the details and potential impacts of a complex project before its implementation, thereby heading off risks before they become too costly [or completely impossible] to correct”.⁴⁸

AIAs, like impact assessments in general, are strongly influenced by the precautionary principle. This principle dictates that, even in the presence of scientific uncertainty regarding the potentially harmful effects of a particular activity, precautionary measures should be implemented⁴⁹ to manage risks, demonstrate the safety of the activity, and prevent serious and irreversible damage. Impact assessments enable us to rectify issues such as using databases that do not accurately represent the social context in which the technology will be applied, which could lead to bias. These assessments also allow for the development of plans to mitigate damage in advance, in case any potentially harmful events occur.

However, although the immediate aim of a prior assessment is to prevent risks, the aim is also to allow the needs and particularities of the social and institutional context in which the system will be introduced to be considered before the system is developed or implemented. It follows that, especially when carried out by the public administration, the aim of an AIA is not only to mitigate risks but to stimulate prior reflection on what types of technologies are best suited to the public interest and how to ensure that their development and use is supported by ethical principles.

In this sense, the precautionary principle mentioned above has also a “connotation of public deliberation”,⁵⁰ encouraging a reasoned and participatory discussion about whether to adopt an algorithmic policy, depending on the types of risks that underlie it. An impact assessment could, for example, reveal that the use of a facial recognition system for public security would inevitably increase police persecution towards black and poor people, leading the public agency to decide not to adopt it in that location.

As AIAs are designed to be carried out before the implementation or even before the development of a given AI system, they are considered an instrument of *ex-ante* accountability.⁵¹ For a prior study of the impacts to be viable, AIAs must be incorporated into the institutional processes through which algorithms are developed and/or acquired.

⁴⁸ Selbst (2021), p. 122.

⁴⁹ Reisman et al. (2018), p. 15.

⁵⁰ See Bioni and Luciano (2020).

⁵¹ See Wieringa (2020).

Public administrations can acquire these systems in various ways, either by developing them in-house, through public purchases, government subsidies, funding from private foundations, and donations. If the systems are sold by technology companies, impact assessments should ideally be incorporated into the existing public procurement or tendering processes.⁵² It should also be noted that although AIAs are designed to take place before the implementation of a given system, ideally, they should also be carried out during its use, and updated on an ongoing basis.⁵³

Emphasis should be placed on the types of *impacts* to be assessed by the AIA, which should not be limited to individual impacts⁵⁴ but must necessarily also encompass social⁵⁵ and collective impacts. This implies considering both (i) *harm to specific groups*, such as black people, women, and LGBTQIA+, and (ii) *harm to democratic rule of law processes* themselves, such as to the principles of transparency, justification, and deliberation required for public decisions.

The first type of impact could, for example, be a situation in which the algorithm disproportionately affects black populations, as has been seen repeatedly in facial recognition and predictive policing systems, used for criminal prosecution purposes. In cases like these, the task of assessing impacts also becomes a means of *tracking down the biases and prejudices* recurrently embedded in the operation of these systems.

In the second type, the impacts that should be considered are related to the democratic rule of law processes per se. It could be a situation in which the introduction of the algorithm into the decision-making process ends up obscuring the data that informs government decision-making and, consequently, preventing society from being able to understand its outcome or even assess whether the decision or policy implemented is lawful or compatible with the public interest.

Ideally, this process of identifying and assessing impacts should be carried out with the participation of various actors and should be the result of multiple perspectives, especially the perspectives of the communities affected by these systems. The definition of *which* impacts will be considered—and potentially addressed—depends directly on *who* participates in the impact assessment process.

It is in this sense that Moss et al. (2021) conceptualize impacts as “boundary objects” since they reflect disputes between the various economic, political, and social interests that are at stake when it comes to implementing these technologies. As the authors point out:

[...] the process of identifying, measuring, formalizing, and accounting for ‘impacts’ [of an algorithmic intervention] is a power-laden process that does not have a neutral endpoint.

⁵² Reisman et al. (2018), p. 8.

⁵³ See Reisman et al. (2018).

⁵⁴ As conceptualized by Smuha (2021), individual harm caused by AI systems “occurs when one or more interests of an individual are unduly violated”. That is, “at the receiving end of the harm is an identifiable individual” (p. 5).

⁵⁵ As pointed out by Smuha (2021), social damage is a *sui generis* type of damage, which “occurs when one or more of society’s interests are unjustly violated”, such as social equality, the integrity of democratic processes and the rule of law.

Precisely because these systems are complex and multi-causal, defining what counts as an impact is contested, shaped by social, economic, and political power.⁵⁶

In addition to a preventive function, the risk assessment made possible by AIA can also have a regulatory function. This is the case with the impact assessment implemented by Canada, for example. As will be approached later in this contribution, the Canadian impact assessment system relies on an electronic questionnaire to provide a score for the AI system. This score helps in determining the level of risk associated with the system. Depending on the classification provided by the AIA (such as excessive, high, and moderate risk), specific regulatory obligations applicable to the case can be defined. As stated by Selbst, the score provided by the Canadian AIA “determines whether the overall risk falls within one of four wide bands (Impact Levels I–IV), and agencies implementing algorithmic systems that fall within a given band must take certain increasingly involved remedial actions to mitigate the anticipated harms”.⁵⁷

Classifying the system's degree of risk to assign the appropriate regulatory measures is a strategy of risk-based regulation, that is being increasingly adopted not only by the Canadian regulation,⁵⁸ but also by the AI Act approved in March 2024 by the European Parliament, and by ongoing proposals, such as the Legal Framework for Artificial Intelligence that is currently being debated in Brazil.⁵⁹ However, unlike Canada, in the European and Brazilian regulatory initiatives, the degree of system risk is not determined directly by impact assessments, but rather by preliminary assessments, conducted before algorithmic impact assessments.

3.3 *Weighting of Costs and Benefits*

The previous identification of impacts and risks contributes to the ability to decide whether implementing an algorithmic system is appropriate, whether it should be done or should undergo adjustments. This is linked to a third purpose of impact assessment, which is the *weighting of costs and benefits*.

Weighting consists of analyzing whether the project to be implemented will be effective in solving the problem, whether the benefits outweigh the costs—including economic costs—of implementing the measure, and whether the proposed policy is the best alternative for dealing with the issue at hand. Weighing costs and benefits is a crucial activity for the effectiveness of public management, meaning

⁵⁶ Moss et al. (2021), p. 9.

⁵⁷ See Selbst (2021), p. 143.

⁵⁸ Government of Canada (2019). Directive on Automated Decision-Making. <https://www.tbs-sct.canada.ca/pol/doc-eng.aspx?id=32592>.

⁵⁹ The Brazilian Legal Framework for Artificial Intelligence (Law Project 2338/2023) is the main project to regulate AI that is being discussed in Brazil and is currently waiting to be voted in the Federal Senate. Available at: <https://www25.senado.leg.br/web/atividade/materias/-/materia/157233>.

the state's ability to manage its resources and decisions in a way that efficiently serves the public interest.

3.4 Public Participation

According to Reisman and colleagues,⁶⁰ AIA goes beyond an internal study carried out by the body that intends to apply AI. AIAs must necessarily include external input from society. The process must guarantee the opportunity for *meaningful public participation*. This includes the possibility for society to contribute to the definition of impacts and risks through interviews and studies presented in public consultations, to challenge specific information presented in the initial assessment report, or even to disagree with the implementation of the system as a whole.

For Moss et al, AIA should reshape power relations in developing and deploying algorithmic systems.⁶¹ By embracing participatory processes, the AIA aims to be a mechanism not only for preventing harm but also for protecting the public interest. Including a broader public in the discussion aims to prevent the State from implementing algorithmic policies driven by the interests of a ruling elite and/or exclusively by the economic interests of technology companies.

Public participation is an important step for carrying out not just a technical but a socio-technical analysis⁶² of algorithms that address not only their technical failures but also the social consequences of their operations. Social scientists, lawyers, and human rights specialists, for example, have too much to contribute to documenting the potential harms of an algorithmic system in a given social context.

The knowledge that comes from humanities will reveal other aspects and dimensions of the problem than those revealed, for example, by a technical audit that verifies the accuracy of the algorithm. This socio-technical analysis demands the interaction of various fields of knowledge, not only from STEM⁶³ but also from SHAPE⁶⁴ disciplines. Technical and social fields must work together.

Public participation, be it from the people and communities affected by these systems, be it from professionals of many different fields, is a process aiming to guarantee the very effectiveness of the AIA because identifying risks and strategies for mitigating harm requires different types of expertise and multiple points of view.

⁶⁰ See Reisman et al. (2018).

⁶¹ See Moss et al. (2021).

⁶² “Sociotechnical systems require methods that can address their simultaneously technical and social dimensions. The absence of diversity in expertise may fail to capture the entire gamut of impacts. Overly technical assessments with no accounting for human experience are not useful, and vice versa”. See Moss et al. (2021), p. 23.

⁶³ “STEM is an acronym for the disciplines of Science, Technology, Engineering, and Mathematics”.

⁶⁴ SHAPE is an acronym for the disciplines of Social Sciences, Humanities, and the Arts for People, and the Economy.

3.5 Accountability

Besides an instrument for society's participation in developing algorithmic policies, the AIA also functions as an instrument for *governmental accountability*. In other words, public bodies must provide *reasons that justify* their assessment of the algorithmic system (its potential risks and benefits) and the choices arising from this assessment, such as the choice to implement the system. In order to do this, the AIA must “create and provide documentation of the decisions made during development and their rationales, which in turn can lead to better accountability for those decisions (...)”.⁶⁵ As stated by Selbst, AIA is a tool to disclose “how policy goals are translated into algorithmic systems, or the political choices that the algorithmic systems actually represent”.⁶⁶

Therefore, *justification* occurs when the impact assessment “makes visible the subjective choices [of the agencies] responsible for deploying an algorithmic system, including how they define harms, benefits and risks, and how they evaluate and choose between alternative policy options related to the deployment of an algorithmic system”.⁶⁷ In any case, it should be made clear what “actions have been taken to assess and manage the risks at stake”.⁶⁸

In addition to being an instrument for *justifying conduct*, the impact assessment must also be an instrument for *responsibilization*. In other words, the body must take responsibility for the results arising from the impact assessment and the potential implementation of the algorithmic system in the social context. In this sense, some questions to be answered by the impact assessment are: How can the damage resulting from these impacts be mitigated? Who is responsible for repairing these problems? As indicated earlier, a damage mitigation plan must be established, indicating the responsible actors and the duties and measures to be taken in case of potential damage.

4 Criteria and Stages of an Algorithmic Impact Assessment

Impact Assessments are not a new regulatory mechanism, as they have been used in various fields to evaluate the risks, costs, and benefits of a project that could potentially affect the rights and legal goods protected by a legal system.⁶⁹ For instance, Environmental Impact Assessments⁷⁰ are established in many juridical systems as mandatory mechanisms for public and private entities that plan to undertake

⁶⁵ Selbst (2021), p. 122.

⁶⁶ Selbst (2021), p. 121.

⁶⁷ Ada Lovelace et al (2021), p. 22.

⁶⁸ See Bioni and Luciano (2020).

⁶⁹ See Frajthof (2022).

⁷⁰ The USA was a pioneer in regulating Environmental Impact Assessment with the enactment of the National Environmental Policy Act (NEPA) in 1968, which served as a model for several other countries.

activities that may harm or pollute the environment. Similarly, Data Protection Impact Assessments are necessary for companies and public entities that collect personal data, as they help to identify potential risks to the rights of data subjects.⁷¹

In addition to serving as a strategy to regulate risky activities, impact assessments are also instruments that states use to guide their decision-making processes and the formulation of public policies, as is the case with Legislative Impact Assessments, for example. Legislative Impact Assessments have a consistent history of application in countries such as the United Kingdom, Portugal, and Canada, which adopt this process to measure the socioeconomic effects of bills and public policies, assess whether they are supported by concrete evidence, and weigh their necessity and adequacy.⁷²

In general, impact assessments are mechanisms associated with the idea of “*better regulation*” and “*open government*” since, as indicated in the European Commission’s Better Regulation Guidelines, they are “a way of working to ensure that policy decisions are prepared in an open and transparent manner, informed by the best available evidence and supported by the comprehensive involvement of stakeholders”.⁷³ Because they are applied in various fields and regulated in different places worldwide, impact assessments are a field of study, with best practices and principles shared between their various application domains.⁷⁴

Based on a literature review and the lessons learned from existing impact assessment models,⁷⁵ some criteria can be identified for algorithmic impact assessments to be properly institutionalized and incorporated into the routine of the public bodies subject to its application. Among them are establishing (i) binding rules; (ii) a supervisory authority; (iii) a proper scope of application; (iv) and a proper implementation process.

4.1 Binding Rules

For impact assessments to be effective in practice, they must be properly *institutionalized*, usually through a law or internal regulations, establishing the validity and basic parameters for conducting this process.⁷⁶ In general, an impact assessment

⁷¹ DPIAs are required by many legislations worldwide, such as the Data Protection Regulations in the EU and in Brazil. See Frajthof (2022).

⁷² See Meneguín and Silva (2017).

⁷³ See European Commission (2015), p. 5.

⁷⁴ In this regard, see the recommendations of the International Association of Impact Assessment: <https://iaia.org/best-practice.php>.

⁷⁵ A report published by the Data & Society Institute (Moss et al. 2021) identified 10 elements in common between existing impact assessments in various fields. The criteria proposed here were inspired by the elements analyzed in the aforementioned research, as well as other works. See: Moss et al. (2021), Reisman et al. (2018), Ada Lovelace Institute et al. (2021).

⁷⁶ These are what Moss et al. (2021) call the “sources of legitimacy” of impact evaluations.

regime is created by a government body with normative power, which will outline general rules for conducting this process, which may be the legislative branch and/or a regulatory agency.

Attempts to institutionalize algorithmic impact assessments have been gradually implemented in some contexts through various political instruments, from general regulations such as the *AI Act*,⁷⁷ to federal laws, such as the *Directive on Automated Decision-Making*, approved in 2019 in Canada,⁷⁸ to non-binding instruments, such as the *Algorithm Charter for Aotearoa New Zealand*,⁷⁹ recommended by the New Zealand government. Another example is the *Fundamental Rights and Algorithms Impact Assessments (FRAIA)*,⁸⁰ commissioned by the Ministry of the Interior of the Dutch Government,⁸¹ which is the object of a parliamentary motion⁸² asking the government for the instrument to become mandatory in the Netherlands.

AIA is also established in municipal ordinances, as in the case of Seattle⁸³ and New York.⁸⁴ Besides those that are already in force, AIAs are being established in ongoing proposals such as the Brazilian Legal Framework for Artificial Intelligence, which is currently waiting for approval by the Country's National Parliament.⁸⁵

Regardless of which legal instrument is chosen for implementing the rules of AIA, it is important to mention that the standards applicable to algorithmic impact assessments have a lot to learn from the pitfalls and successes of impact assessment models that already exist in other areas. In this sense, some authors⁸⁶ point out that the structure of DPIAs, currently required by foreign and national regulations, such as the European (GDPR) and Brazilian (LGPD) Data Protection Laws,⁸⁷ could serve as a basis for building Algorithmic Impact Assessments.

⁷⁷ The text of the law is available at: <https://artificialintelligenceact.eu/ai-act-explorer/>.

⁷⁸ Government of Canada. (2019). Directive on Automated Decision-Making. <https://www.tbs-sct.canada.ca/pol/doc-eng.aspx?id=32592>.

⁷⁹ Initiative of the New Zealand Ministry of Statistics, which aims to outline general parameters of transparency and accountability for the use of algorithms by public bodies. Available at: <https://data.govt.nz/toolkit/data-ethics/government-algorithm-transparency-and-accountability/>.

⁸⁰ See Gerards et al. (2022).

⁸¹ FRAIA is an impact assessment model that aims to support decision-making, primarily for the public sector, about the development and deployment of AI systems, with a consideration of its impacts on fundamental rights. Available at: <https://www.government.nl/documents/reports/2022/03/31/impact-assessment-fundamental-rights-and-algorithms>.

⁸² Available at: <https://www.tweedekamer.nl/kamerstukken/moties/detail?id=2022Z06024&di d=2022D12329>.

⁸³ Available at: <http://seattle.legistar.com/ViewReport.aspx?M=R&N=Text&GID=393&ID=2849012&GUID=5B7D2F80-A918-4931-9E2E-88E27478A89E&Title=Legislation+Text>.

⁸⁴ Available at: <https://legistar.council.nyc.gov/LegislationDetail.aspx?ID=3343878&GUID=996ABB2A-9F4C-4A32-B081-D6F24AB954A0>.

⁸⁵ The text of the law is available at: <https://www25.senado.leg.br/web/atividade/materias/-/materia/157233>.

⁸⁶ See Kaminski (2019) and Frajhof (2022).

⁸⁷ Personal Data Protection Impact reports (DPIAs) are defined in art. 5, inc. XVII of the LGPD as “documentation from the controller that contains a description of the personal data processing

It is essential to mention that the process of institutionalizing impact assessments depends not only on *binding norms* but also on the *administrative capacity* to enforce and implement these norms into the everyday routine of institutions. Administrative and financial resources and an institutional culture of accountability are critical to ensure that impact assessments go beyond a “tenet of best practice”,⁸⁸ producing significant effects in the real world.

4.2 Supervisory Authority

In addition to binding rules, the institutionalization of AIAs also requires a *supervisory authority*, implemented by the State, with the power to verify compliance with AIA regulations and oversee the actor responsible for preparing the impact assessment.

As previously mentioned, an *external forum* to oversee the accountable actor is an essential component of any accountability mechanism.⁸⁹ Thus, for an algorithmic impact assessment to be an effective instrument for public accountability, the existence of a supervisory authority, playing the role of an external forum, together with civil society, is essential.

This authority must hold investigative and corrective powers to supervise AIA's procedure and judge the adequacy of the information and justifications provided by the entity responsible for formulating the assessment. Moreover, the supervisory authority may also be responsible for regulating impact assessment procedures and assigning consequences if the assessment does not comply with the previously determined parameters.⁹⁰

Data Protection Authorities, already established by many jurisdictions, are likely to assume these functions, due to the competencies these bodies already have in supervising the application of data protection laws and DPIAs. This will probably be the case of Brazil, once the draft bill being currently discussed in the Parliament appoints the National Data Protection Authority as the central authority for the Brazilian National System of AI Regulation.⁹¹

In contrast, in the European scenario, the supervision of impact assessments will be conducted by the Market Surveillance Authorities⁹² designated by each member

processes that may generate risks to civil liberties and fundamental rights, as well as measures, safeguards and risk mitigation mechanisms”. In the GDPR, DPIAs are regulated in Articles 35 and 36.

⁸⁸ See Moss et al. (2021), p. 17.

⁸⁹ See Bovens (2007).

⁹⁰ See Moss et al. (2021).

⁹¹ See article 44 of the Draft Bill n° 2338.

⁹² Article 3 (26) of the AI Act: “‘market surveillance authority’ means the national authority carrying out the activities and taking the measures pursuant to Regulation (EU) 2019/1020”.

state.⁹³ In the case of institutions, bodies, offices or agencies belonging to the European Union, the corresponding market surveillance authority should be the European Data Protection Supervisor.⁹⁴ As provided by the AI Act, the market surveillance authorities should be notified of the results of the impact assessments as soon as they have been carried out by AI deployers.⁹⁵ A very similar provision is also encountered in the Brazilian proposal, determining that “developers of artificial intelligence systems must share preliminary and algorithmic impact assessments with the competent authorities, under the terms of the regulation”.⁹⁶

Due to the regulatory, supervisory, and sanctioning powers to be performed by a supervisory authority, its importance for the effectiveness of impact assessments is critical. Nonetheless, there are concerns that impact assessments become merely formal requirements to be fulfilled by the agencies that aim to apply an algorithmic system without playing an influential role in accountability and risk management.⁹⁷ The supervisory authority's performance in regulating and overseeing AIAs is decisive for their effectiveness and should be the subject of attention in current and future regulations.

4.3 Scope of Application

Another necessary element for a proper institutionalization of AIAs is establishing its *scope of application*. What types of AI systems should be subject to an impact assessment? Although there are still no categorical guidelines in this regard, delimiting the scope of AIAs is key to the success of this instrument.⁹⁸

In the city of Seattle, USA, for example, municipal departments are required to prepare “surveillance impact reports” specifically for the deployment of algorithmic

⁹³Article 70 (1) of the AI Act: “1. *Each Member State shall establish or designate as national competent authorities at least one notifying authority and at least one market surveillance authority for the purposes of this Regulation. Those national competent authorities shall exercise their powers independently, impartially and without bias so as to safeguard the objectivity of their activities and tasks, and to ensure the application and implementation of this Regulation. The members of those authorities shall refrain from any action incompatible with their duties. Provided that those principles are observed, such activities and tasks may be performed by one or more designated authorities, in accordance with the organizational needs of the Member State*”.

⁹⁴Article 70 (9) of the AI Act: “Where Union institutions, bodies, offices or agencies fall within the scope of this Regulation, the European Data Protection Supervisor shall act as the competent authority for their supervision”.

⁹⁵Article 27: “3. Once the assessment referred to in paragraph 1 of this Article has been performed, *the deployer shall notify the market surveillance authority of its results*, submitting the filled-out template referred to in paragraph 5 of this Article as part of the notification. In the case referred to in Article 46(1), deployers may be exempt from that obligation to notify”.

⁹⁶Article 25 § 1° of the Brazilian Law Proposal 2338/2023.

⁹⁷See Moss et al. (2021) and Reisman et al. (2018).

⁹⁸See Ada Lovelace Institute et al. (2021) and Moss et al. (2021).

surveillance systems. The municipal ordinance⁹⁹ states that the impact reports must contain, among other information, a description of the potential impacts of the surveillance technology on civil rights and liberties, especially of black and marginalized populations; a damage mitigation plan and a fiscal impact assessment, with an indication of expenses and potential cost savings to be generated by the implementation of the technology.

In the case above, an impact assessment is essential, given surveillance systems' influence on fundamental rights. However, algorithmic systems can be applied in public management for the most diverse purposes—be it for internal administrative activities, be it for the provision of public services and law enforcement.¹⁰⁰ Due to the variety of purposes for which algorithmic systems are used, it is necessary to understand in which cases an impact assessment is necessary.

Interviews conducted by the Ada Lovelace Institute et al with public officials from different locations revealed that a common concern regarding algorithmic impact assessments is “the risk of overburdening public agencies by placing mundane or routine algorithmic processes which do not appear to have significant social impacts under review”.¹⁰¹

To maintain administrative efficiency, a broad obligation to carry out impact assessments for every algorithmic system implemented by the public administration may not be viable. On the other hand, broad exemptions could also cause public institutions to shirk their responsibility to carry out impact assessments, especially in cases where they are most needed. For this reason, regulations need to set parameters.

One parameter that can be adopted to determine which systems must undergo an algorithmic impact assessment is the degree of risk they present.¹⁰² Both the AI Act and the Brazilian Legal Framework for Artificial Intelligence determine that AIAs are mandatory for high-risk systems - those that pose an elevated degree of harm to people's health, safety, or fundamental rights.

Some examples of high-risk systems provided both by the AI Act and the Brazilian bill are: (i) those used for biometric identification and authentication; (ii) those used to evaluate the access and eligibility of natural persons to essential private services, such as access to credit, health and life insurances; (iii) those used for educational and professional admissions and assessments, such as access to educational institutions or recruitment of job candidates; (iv) those that impact access to essential public services like health, social assistance, and security; (v) those intended to be used as safety components in critical infrastructures, such as in road traffic and in the supply of water, gas, heating or electricity; (vi) those used by law

⁹⁹ Ordinance 123576 of the City of Seattle. Available at: <http://seattle.legistar.com/ViewReport.aspx?M=R&N=Text&GID=393&ID=2849012&GUID=5B7D2F80-A918-4931-9E2E-88E27478A89E&Title=Legislation+Text>.

¹⁰⁰ The classification of algorithms used for internal administration, provision of public services, law enforcement and awarding of benefits is elaborated by Fernández (2022).

¹⁰¹ Ada Lovelace Institute et al. (2021), p. 41.

¹⁰² See Lemos et al. (2023).

enforcement authorities for profiling natural persons during the detection, investigation, or prosecution of criminal offenses in order to assess the risk of an infraction; (vii) those used for managing immigration and border control; (viii) those used by judicial authorities to investigate facts and apply the law.¹⁰³

From an analysis of the high-risk systems list indicated by both European and Brazilian regulatory initiatives, it is prominent that a significant portion of the systems identified as high-risk are those applied by the public sector. More specifically, by the public administration, for law enforcement and public services provision. This reinforces the argument that impact assessments will be increasingly demanded, not only from private actors but especially from public agents, who will be responsible for implementing a huge part of high-risk AI systems.

Concerning the AI Act, it emphasizes the duty of public actors and providers of public services to carry out impact assessments. According to Article 27 of the AI Act, the conduction of Fundamental Rights Impact Assessments (FRIA) for high-risk systems is mandatory for “bodies governed by public law” and for “private entities providing public services”. Within the context of the private sector, a FRIA is also required for “deployers of high-risk AI systems” used for the evaluation of the creditworthiness of natural persons, to establish their credit score, or for assessing risks and pricings in the context of life and health insurance.¹⁰⁴

There’s, however, an exception, in which even when implementing a high-risk system, a FRIA won’t be needed. This is the case of AI systems “intended to be used as safety components in the management and operation of critical digital infrastructure, road traffic, or in the supply of water, gas, heating or electricity”.¹⁰⁵ As pointed out by Article 27 of the AI Act, the deployers of these systems are exempt from conducting a FRIA.

To identify the obligation to conduct an impact assessment, it is first necessary to know whether the system is high risk. The classification of the degree of risk of the system, as stated by the AI Act, must be done through conformity assessments conducted by the provider or by a third party, as prescribed by the Law.¹⁰⁶

Similarly to the AI Act, the Brazilian regulatory proposal also demands a conformity assessment, named “preliminary assessment”, for establishing the AI system’s degree of risk, followed by an AIA, when it comes to high-risk systems. The law proposal states that before using or introducing an AI system on the market, “the developer and deployer of artificial intelligence systems must carry out a preliminary assessment that will determine their degree of risk”.¹⁰⁷ If the system is

¹⁰³ Annex III of the AI Act and Article 14 of the Brazilian Bill.

¹⁰⁴ Article 27 of the AI Act.

¹⁰⁵ Annex III (2) of the AI Act.

¹⁰⁶ Article 43 of the AI Act.

¹⁰⁷ Article 12 of the Brazilian Legal Framework for Artificial Intelligence (Law Proposal 2338/2023). Available at: https://legis.senado.leg.br/sdleg-getter/documento?dm=9640105&ts=1718747084050&rendition_principal=S&disposition=inline.

classified as high-risk¹⁰⁸ by the preliminary assessment, it must necessarily undergo an algorithmic impact assessment.

Thus, in general, according to the AI Act and the Brazilian bill, systems initially classified as high-risk must be submitted to an algorithmic impact assessment, while other systems would, in principle, be exempt from this procedure. It seems to be a reasonable requirement, aimed, on the one hand, at preventing harm to individuals and democratic processes and, on the other, at not burdening AI deployers with excessive regulatory obligations for systems that do not pose a significant risk. Furthermore, the requirement of AIAs specifically for high-risk systems stems from a central logic of risk-based regulations, whose key feature is the establishment of asymmetric obligations based on the system's degree of risk.

4.4 *Implementation Process*

There is no fixed model for algorithmic impact assessments, which may vary depending on the application context.¹⁰⁹ Nonetheless, it is possible to identify some essential steps that should be part of AIA processes. Based on the impact assessment processes that already exist in other domains and the specific recommendations that have been developed by the academic literature in the field of algorithmic accountability and AIA, it is possible to identify that the algorithmic impact assessment process in the public administration should consist of four main stages: (a) self-assessment; (b) information disclosure; (c) public consultation and (d) contestation.

4.5 *Self-evaluation*

The first stage consists of the impact analysis itself, which should be the responsibility of the public agency planning to implement a given AI system. As recommended by the report produced by the AI Now Institute (2018), government agencies should “perform a self-assessment of each system, evaluate potential issues of inaccuracy, bias, and harms to affected communities, and establish ways to address these potential impacts, including proactive conversations or engagement with affected community members”.¹¹⁰

As noted above, both European and Brazilian regulatory initiatives determine that public and private entities that develop or deploy high-risk systems must

¹⁰⁸Article 14 of the Law Proposal 2338/2023 lists which systems are considered high-risk, according to the purpose for which they are used.

¹⁰⁹See Lemos et al. (2023).

¹¹⁰Reisman et al. (2018), p. 9.

conduct an AIA before the system's use or introduction in the market. The Brazilian law proposal states that AIA must be conducted by a "professional or team of professionals with the necessary technical, scientific, regulatory, and legal expertise, considering sectorial best practices and international standards".¹¹¹

It's worth noting that the professionals responsible for conducting the AIA can be either internal professionals, who are part of the organization that develops or deploys the system, or external professionals, who, in some cases, will also be responsible for conducting an auditing of the AIA carried out by the AI agents. As provided by the law proposal,¹¹² the competent authority must regulate, in the future, the cases in which AIA should be developed by external professionals and also the cases in which external audits on the AIA previously conducted by the agents will be needed.

Conversely, in the European context, the AI Act establishes that FRIAs must be conducted by AI deployers.¹¹³ Nonetheless, providers also have a role to play in the assessment process. As they are responsible for conducting conformity assessments, it is expected that the information produced at this stage should be used by deployers in the FRIA's process. As stated by Mantelero, the conformity assessment and the FRIA are linked, and although the AI Act is not clear about the provider's obligation to share the information from its assessment with the deployer, one assessment should complement the other for an "integrated risk management".¹¹⁴ This does seem to be the purpose of the AI Act, which stipulates that "the deployer may [...] rely on previously conducted fundamental rights impact assessments or existing impact assessments carried out by the provider".¹¹⁵

To facilitate the conduction of FRIAs by AI deployers, the Act indicates that it is the role of the AI Office to "develop a template for a questionnaire, including through an automated tool, to facilitate deployers in complying with their obligations [...] in a simplified manner".¹¹⁶ Initiatives that can facilitate compliance with regulatory obligations are welcome. However, utilizing prefabricated questionnaires to prepare impact assessments raises concerns about their ability to effectively identify, prevent, and mitigate the specific risks presented in each case.

To properly assess the risks, it is necessary to understand that AI technologies are socio-technical systems. They transform the environment and are also transformed by the environment in which they are inserted. Consequently, their risks cannot be assessed solely based on a formal analysis of the technology's technical design, but rather on how the technology will interact with the social variables of each environment.

¹¹¹ See article 26 of the Draft Bill 2338/2023.

¹¹² See Article 26, § 1° of the Draft Bill 2338/2023.

¹¹³ See Article 27 of the AI Act.

¹¹⁴ Mantelero (2024), p. 6.

¹¹⁵ See Article 27 (2) of the AI Act.

¹¹⁶ See Article 27 (5) of the AI Act.

For that reason, the risks these systems may pose to fundamental rights can't be assessed abstractedly, from a simple checklist, but from an in-depth analysis of the social and institutional context in which the system will be implemented. As also noted by Mantelero, "questionnaires or templates to assist deployers can be useful, but without reducing the FRIA to a mere questionnaire-based exercise, which is contrary to its nature".¹¹⁷

At the time of this writing, there are no specifications regarding the structure of the questionnaire that will be developed by the AI Office, nor on the timeline for its implementation. In any event, the government of Canada gives us an example of an AIA that is based solely on an electronic questionnaire.

The referred instrument was implemented in Canada in 2019 by the Federal Directive on Automated Decision Making, which requires any company wishing to contract with the government and any public body wishing to implement these systems to carry out an AIA.¹¹⁸ To this end, the government has made available on its website a questionnaire¹¹⁹ consisting of approximately 80 questions,¹²⁰ to be answered by the actors obliged to carry out an AIA.

Based on the answers provided regarding the data, design, and type of decision generated by the algorithm, the Canadian electronic questionnaire assigns a score to the algorithmic system, that indicates its level of risk. As noted by Moss et al, this questionnaire is a self-assessment mostly based on "yes" or "no" questions, which takes around 35 min to complete. Because it is based on a closed questionnaire, this impact report doesn't allow justification on the answers given, nor the provision of evidence on how the impacts were assessed by the actors.¹²¹

As already criticized by Moss et al, these types of impact assessment "do not foster engagement with real-world effects or the assessment of novel harms",¹²² beyond those already specified in the forms. Moreover, as the possible impacts are already indicated by the form, the process does not enable the participation of different actors and forms of expertise in defining the impacts, as desirable. In cases like this, the purpose of AIA seems to be reduced to categorizing and documenting the degree of risks of an algorithmic system.

For this reason, the impact assessment model implemented in Canada is more similar to a risk classification mechanism, such as the "preliminary assessments" of

¹¹⁷ Mantelero (2024), p. 7.

¹¹⁸ See Government of Canada (2019). Directive on Automated Decision-Making. <https://www.tbs-sct.canada.ca/pol/doc-eng.aspx?id=32592>.

¹¹⁹ Available at: <https://canada-ca.github.io/aia-eia-js/>.

¹²⁰ As indicated by one of the Treasury Board of Canada consultants responsible for developing the algorithmic impact assessment process, questions such as: Are the impacts of the decisions generated by the algorithm reversible? Who is responsible for maintaining, developing and improving the system? How explainable are the machine's determinations? Are you monitoring unintended results? Do you have a clear audit trail around decisions? Do you have an appeal mechanism in place in case people have questions or complaints? See Karlin (2018).

¹²¹ See Moss et al. (2021).

¹²² Moss et al. (2021), p. 23.

the Brazilian proposal, rather than in-depth impact assessments. For that matter, it is important to distinguish between assessments that simply indicate a system's degree of risk through a questionnaire, and more substantial impact assessments, such as the one recommended in this contribution.

In this regard, Selbst makes a distinction between three categories of AIA proposals: “(1) models based on NEPA, (2) models based on the GDPR's DPIA, and (3) a questionnaire model”.¹²³ In this respect, it is worth noting that the type of AIA recommended by this contribution aligns with what Selbst calls “NEPA-based models”. These models are inspired by the Environmental Impact Assessments established by the National Environmental Policy Act (NEPA) in the USA, which served as a model for impact assessments in several other countries.

As stated by Selbst, NEPA-based models imply a “highly detailed impact assessment [that encompasses] thorough answers to open-ended questions that explain the design process” of the AI systems.¹²⁴ These assessments also demand a more advanced level of transparency and public participation.

Considering the NEPA-based model and the recommendations provided by the literature on AIA for the public sector,¹²⁵ the goal of AIA is not only to determine the degree of risk of an algorithm but also to evaluate its potential impacts on fundamental rights. Especially in the public administration context, the role of AIA is to foster reflection and public debate on *whether, how and under what conditions* the implementation of an AI system is desirable.

Therefore, the self-assessment carried out by the public body that intends to implement an algorithmic system must include information on its potential impacts, the costs and benefits of the measure, and a mitigation plan for eventual harm. Given the importance of a contextual and in-depth analysis of the risks and impacts of AI systems on fundamental rights, some studies, such as those by Mantelero¹²⁶ and Lemos and colleagues¹²⁷ seek to propose methodologies that can guide the formulation of impact assessments in Europe and in Brazil, respectively.

4.6 Disclosure of Information

Once the self-assessment is completed, it is time to disclose information produced during this analysis. Publishing the impact assessment is widely recommended by studies in the field¹²⁸ and is already mandated by some regulatory initiatives, such as the Directive on Automated Decision-Making in Canada and the Framework for AI

¹²³ Selbst (2021), p. 140.

¹²⁴ Selbst (2021), p. 141.

¹²⁵ See Moss et al. (2021), Reisman et al. (2018) and Ada Lovelace et al. (2021).

¹²⁶ See Mantelero (2024).

¹²⁷ See Lemos et al. (2023).

¹²⁸ See Moss et al. (2021), Reisman et al. (2018) and Lemos et al. (2023).

Regulation in Brazil. Disclosing the impact assessment results is especially important for AI systems implemented by the public administration, as they need to be scrutinized by society and by the other government branches.

In view of this, the body responsible for conducting the assessment must publicize the report, with information regarding the intended purposes of the system, implementation schedule, and expected social and economic impacts. This information disclosure stage must occur after the body's self-assessment, and after the completion of the entire algorithmic impact assessment process.¹²⁹

Two formats are recommended for AIA disclosure: a detailed and technical one for expert analysis, and a summarized format, easily accessible to the general public. According to Reisman and colleagues,¹³⁰ the dual disclosure model is a practice already implemented in other impact assessment initiatives, such as in the environmental and data protection domains. Its purpose is to promote comprehension and social control among a broader audience, beyond experts.

The purpose of the information disclosure stage is to provide transparency regarding the use of algorithms by the public administration and guarantee society's right to know how these systems work and affect their rights. Therefore, the main concern at this stage is to specify the impact report's *level of public access*. As indicated in research carried out by the Data & Society Institute (2021):

Every impact assessment process must specify its level of public access, which determines who has access to the impact statement reports, supporting evidence, and procedural elements. [...] The broader the access to its impact statement, the stronger is an impact assessment's potential to enact changes in system design, deployment, and operation.¹³¹

When algorithms are used by public administration, the impact assessment report should be disclosed proactively as it is of public interest. This means it should be accessible without the need for prior request, under a regime of active transparency.¹³²

Publicizing impact reports¹³³ is essential to enable social control over the algorithmic policies implemented by the government, especially considering such information would be difficult to obtain through information access isolated requests.¹³⁴ In addition, making this information public is a precondition for society to be able

¹²⁹ Ibid.

¹³⁰ See Reisman et al. (2018).

¹³¹ Moss et al. (2021), p. 20.

¹³² Active transparency is a requirement of the Brazilian Access to Information Law, as well as a consequence of the publicity principle, which informs state acts not only in Brazil but also in most democracies. On the Brazilian Access to Information Law, see: Barcellos (2015). On the right of access to information in the context of artificial intelligence, see: Keller (2019).

¹³³ The obligation to publicly disclose the impact report is a feature generally present in environmental impact assessment regimes. However, not all impact assessments are originally formulated as publicly accessible. DPIAs and human rights impact assessments (HRIAs) are published at the discretion of the companies that carry them out, which raises doubts about the effectiveness of these processes as accountability instruments.

¹³⁴ Reisman et al. (2018), p. 17.

to participate in the subsequent phase of the impact assessment process: *public consultations*.

The obligation to publicly disclose the impact report is a feature generally present in environmental impact assessment regimes. However, not all impact assessments are originally formulated as publicly accessible. DPIAs and human rights impact assessments (HRIAs) are published at the discretion of the companies that carry them out, which raises doubts about the effectiveness of these processes as accountability instruments.

As with environmental impact assessments, the obligation to publicly disclose algorithmic impact assessments is already established in some current AI regulations. The Canadian *Directive on Automated Decision-Making* (2019), for instance, states that the body using an automated decision system must be responsible for “releasing the final results of the Algorithmic Impact Assessment in an accessible format via Government of Canada websites”.¹³⁵ In Brazil, the aforementioned proposal for AI regulation also provides for both preliminary assessments and algorithmic impact assessments of systems adopted by the public sector to be published in easily accessible communication vehicles.¹³⁶

Unlike the aforementioned initiatives, the AI Act does not provide for the publication of algorithmic impact assessments, but only for market surveillance authorities to be notified of their results.¹³⁷ In order to establish some degree of public transparency about the high-risk systems implemented by the public administration, the AI Act determines that they be registered in the EU database.¹³⁸

Most of the information contained in this database “shall be accessible and publicly available in a user-friendly manner [,] easily navigable and machine-readable”.¹³⁹ It can be observed that the main purpose of the European database is to serve as a public registry for high-risk AI systems, providing a certain level of transparency and societal oversight over these systems. Despite demands from civil society to deposit the FRIA’s output into public registers,¹⁴⁰ this requirement was not included in the approved version of the Act.

Besides the AI Act, public disclosure of information about algorithms implemented by the state is already being made by other existing initiatives. In the US, for example, it has already been mandatory since 2021 for federal government agencies to create a public inventory of how they use AI.¹⁴¹ The introduction of public *AI Registries* has also become a requirement of the municipal administration of some

¹³⁵ Government of Canada (2019). Directive on Automated Decision-Making. <https://www.tbs-sct.canada.ca/pol/doc-eng.aspx?id=32592>.

¹³⁶ Article 31 of the Draft Bill 2338/2023.

¹³⁷ Article 27 (3) of the AI Act.

¹³⁸ Articles 26, 49 and 71 of the AI Act.

¹³⁹ Article 71(4) of the AI Act.

¹⁴⁰ Data & Society and ECNL, p. 3.

¹⁴¹ This advertising rule came into force with Executive Order 13960, published in Dec. 2020. In this regard, see: <https://www.federalregister.gov/documents/2020/12/08/2020-27065/promoting-the-use-of-trustworthy-artificial-intelligence-in-the-federal-government>.

cities, such as Amsterdam, Helsinki, and Seattle¹⁴² that already make publicly available information on the AI systems in force on their official websites.

Despite its importance, disclosure of information on algorithms is not always a simple task. One of the challenges administrations may face at this stage is the allegation of violating commercial secrecy, often brought when the technology is developed by private companies. It is the role of public agencies to analyze, before entering into a contract with technology suppliers, whether trade secret claims by the company could frustrate the right to meaningful access to information.

As indicated by Reisman and colleagues, “vendors should be contractually required by [public] agencies to waive any proprietary or trade secrecy interest in information related to accountability, such as those surrounding testing, validation, and/or verification of system performance and disparate impact”.¹⁴³ Where trade secrets exist, the contractor's confidentiality rules should be indicated in the impact assessment, but it should not eliminate the public agency's responsibility to be accountable to society.

The *information disclosure stage* is crucial for transparency of state actions and social control by civil society. This stage allows citizens, researchers, external auditors, and non-governmental organizations to analyze the information provided by the initial report and prepare contributions to be presented at the subsequent *public consultation stage*.

4.7 Public Consultation

Public consultations are a mandatory stage in Environmental Impact Assessment processes as they should also be for Algorithmic Impact Assessments. As widely recommended in academic literature, public consultations are fundamental to AIA,¹⁴⁴ especially those conducted by the public administration. It is during the public consultation phase that interested parties can comment on the content of the impact report and provide or request information, research, and evidence that complement the initial assessment.¹⁴⁵ The public's contributions can highlight impacts and risks that haven't been previously considered by the initial assessment.

Ideally, public input can influence the state's decision on whether to adopt the algorithmic system, as well as promote possible reforms to the policy. AIAs public consultation can work as a process of “evidence gathering”,¹⁴⁶ in which interviews are conducted to investigate the possible impacts of the policy, from the perspective

¹⁴² See Floridi (2021) and Haataja (2020).

¹⁴³ Reisman et al. (2018), p. 15.

¹⁴⁴ See Reisman et al. (2018), Moss et al. (2021), Ada Lovelace et al. (2021), Selbst (2021), Lemos et al. (2023) and Mantelero (2024).

¹⁴⁵ Reisman et al. (2018), p. 9.

¹⁴⁶ Moss et al. (2021), p. 21.

of experts, individuals, and affected communities.¹⁴⁷ As indicated by Moss et al, it is important to consider: (i) who are the stakeholders? (ii) how should the relationship between these actors be? and (iii) how different forms of expertise are represented in these participation processes?¹⁴⁸

The actors who must participate in AIA public consultations can vary. In principle, the more diverse the stakeholders participating in the process, the higher the chances for the impact assessment to address different types of risks. Therefore, an AIA can count on the participation of system developers, representatives of the technology companies responsible for manufacturing them, public officials from the government agencies that aim to develop, buy, or license these systems, research institutes, scientists, interested citizens, independent auditors who can conduct technical investigations, civic organizations working on digital rights and privacy-related issues, local organizations defending communities rights, and government agencies in charge of supervision and regulation.

Despite the importance of the participation of different actors, there's still a "lack of methods for deciding what the relevant groups are and who gets to represent them".¹⁴⁹ Moreover, public participation in AIA is likely to face challenges already inherent to traditional participatory procedures, such as economic and social barriers to inclusion, the public's lack of information about their rights, power asymmetries between the political interests at stake, financial and time constraints, and an institutional culture that is generally closed to participatory practices.

In the face of these challenges, there is great concern that public consultations become a "hollow requirement"¹⁵⁰ as is already seen with privacy and data protection impact assessments, where although consultations are formally mandated, there is no specification on their required goals and processes, nor proper enforcement mechanisms.

To avoid these pitfalls, AIA regulations must determine how the public participation process will be organized and how research and audit work can be remunerated.¹⁵¹ Nonetheless, as discussed by Selbst, although "many of the specific AIA proposals envision input from affected communities, none of the proposals has much detail or a workable vision for how to accomplish it, nor a way to go beyond mere input to some form of power over the system".¹⁵²

Fortunately, the Brazilian draft bill establishes the need for social participation in the AIA, determining that the "elaboration of the impact assessment will include, depending on the risk and economic size of the organization, the public

¹⁴⁷This already occurs in fiscal impact assessments and human rights impact assessments, as mentioned by Moss et al. (2021), p. 21.

¹⁴⁸See Moss et al. (2021).

¹⁴⁹See Selbst (2021), p. 161.

¹⁵⁰Moss et al. (2021), p. 21.

¹⁵¹Moss et al. (2021) and Selbst (2021).

¹⁵²Selbst (2021), p. 160.

participation of the different social segments affected, especially vulnerable groups potentially affected by the systems, under the terms of the regulation”.¹⁵³

Despite requiring the participation of society in the preparation of impact assessments, the Brazilian proposal does not specify how this should take place, leaving this procedure to be regulated by the competent authorities in the future. In this regard, the bill also states that “it will be up to the competent authorities to establish the hypotheses in which the public participation [...] will be dispensed with, as well as the hypotheses in which it may be carried out in a simplified manner, indicating the criteria for this participation”.¹⁵⁴

Although social participation is strongly recommended, especially in AIAs conducted by the public administration, it was not demanded by the final version of the AI Act, going against best practices in impact assessment.¹⁵⁵ The only reference made by the AI Act to social participation in FRIA is in Recital 96, which establishes that “deployers of high-risk AI system, in particular when AI systems are used in the public sector, could involve relevant stakeholders, including the representatives of groups of persons likely to be affected by the AI system, independent experts, and civil society organizations [...]”.¹⁵⁶

While drafting the AI Act, several research centers and civil society organizations called for public consultations to be included in the FRIA’s process.¹⁵⁷ As a result of civil society efforts, the previous version of the AI Act, drafted by the European Parliament, determined the necessity of public consultations. According to the previous text, which unfortunately was not approved, the deployer conducting an AIA should “to the best extent possible, involve representatives of the persons or groups of persons that are likely to be affected by the high-risk AI system”.¹⁵⁸

The referred text exemplified some stakeholders that should be notified by the deployer to provide input into the impact assessment, such as “equality bodies, consumer protection agencies, social partners, and data protection agencies”.¹⁵⁹ It is worth noting that, as stated by Mantelero, even the proposed provision “suggested a rather limited form of participation”.¹⁶⁰ However, even this limited form of participation was removed from the final version of the AI Act in the last stages of negotiations.¹⁶¹

Considering the above, the current text of the AI Act does not provide a legally binding way for society to contribute to impact assessments. However, the absence

¹⁵³ Article 28 of the Draft Bill 2338/2023.

¹⁵⁴ Article 28, sole paragraph of the Draft Bill 2338/2023.

¹⁵⁵ Mantelero (2024), p. 8.

¹⁵⁶ Recital 96 of the AI Act.

¹⁵⁷ As reported by the ECNL, more than 150 civil society organizations called EU institutions to meaningfully engage civil society and affected communities in FRIAs. See ECNL (2023).

¹⁵⁸ Mantelero (2024), p. 8, footnote 70.

¹⁵⁹ Mantelero (2024), p. 8, footnote 70.

¹⁶⁰ See Mantelero (2024).

¹⁶¹ Article 19 of the AI Act.

of a mandatory participation procedure in the AI Act does not mean that such an obligation cannot be required by other regimes, such as national laws,¹⁶² and, eventually, by competent authorities.

In any event, it is regrettable that European regulation has neglected the importance of social participation in impact assessments. The lack of mechanisms in the AI Act for society to participate in the identification and prevention of risks raises concerns about the effective protection of fundamental rights¹⁶³ not only in Europe but in countries whose AI regulation is still emerging and is heavily influenced by the Brussels effect.

By neglecting the importance of civil society participation in AI governance processes, regulators miss the opportunity to rely on the expertise and point of view of people who will continue to be constantly affected by these systems and who should therefore have the opportunity to have a say and “a measure of control”¹⁶⁴ on how AI systems influence their lives.

4.8 *Contestation*

In addition to public consultations, some academic literature recommends that AIAs include a contestation stage.¹⁶⁵ Although we couldn’t find evidence of regulatory provisions or practical application of this procedure, a contestation stage should be designed to give opportunities for the public to contest the impact assessment.

This contestation should be performed before the involved administrative authority itself or a higher administrative authority (administrative appeal), a supervisory agency, or even a court, in case of irregularities in the process or if the actor responsible for conducting the AIA refuses to rectify problems pointed out by other stakeholders during the public consultation period.

A recurring concern in the literature is that impact assessments become mere “checklists” designed to provide superficial compliance rather than truly being a mechanism of public debate and risk mitigation. The introduction of a stage for contestation of AIA seeks to remedy this problem. Therefore, it requires legal mechanisms that enable society to report any such irregularities and, in certain cases, to oppose the implementation of algorithmic systems that are not in line with the public interest or that threaten fundamental rights.

One possible procedure at this stage is the request for external audits by independent researchers and institutions to be carried out before the system is implemented. Requesting external audits is recommended in cases where civil society or regulators suspect that the agency’s internal assessment is incomplete or has failed to

¹⁶² See Mantelero (2024), p. 17.

¹⁶³ See Mantelero (2024) and Article 19 of the AI Act.

¹⁶⁴ Selbst (2021), p. 161.

¹⁶⁵ Reisman et al. (2018), p. 10.

consider specific risks and harms. As indicated by Reisman and colleagues, in situations where internal agency assessments are insufficient or where particular risks or harms have gone unaddressed, external researchers and auditors could raise the need for [external] review”¹⁶⁶ before the system’s deployment.

In this respect, the Brazilian draft bill offers an interesting proposal. Although a contestation stage for AIA is not explicitly provided by the Brazilian Framework for AI regulation, the current version of the proposal determines that “the competent authority, together with the sectoral authorities, may carry out joint investigations into high-risk AI systems, in cases of suspected violation of the principles, rights and duties provided for in this Law or in the applicable sectoral legislation”.¹⁶⁷

Moreover, the bill also establishes that the competent authority must “accredit institutions (...) to access data for audit purposes”.¹⁶⁸ It is therefore hoped that audits, carried out either by independent bodies or by regulatory authorities, can function as a form of external verification, or even external challenge of the impact assessment conducted by the deployers, as well as being an extra accountability procedure aimed at guaranteeing the protection of fundamental rights.

Finally, impact assessments should be an iterative and dynamic process. To ensure that AIA is up to date, it is recommended that government agencies regularly renew their processes, especially when changes are made to the system or in its application context. The need to update the impact assessment can also arise due to the evolution of new scientific parameters and recommendations applicable to the case.¹⁶⁹ Thus, a set period may be established for the body to carry out a new assessment, which may consequently require opening up new opportunities for public consultation and contestation.

5 Conclusion: Challenges of AIA as an Accountability Mechanism

As explored throughout this contribution, AIA is an essential mechanism for providing more responsible use of AI by governments. For this to happen, AIA should not be viewed only as a risk management mechanism. On the contrary, it must also be understood as an instrument of reflection and public debate on the adequacy of an AI system in a given social context. Moreover, it must be a mechanism of information and social control over algorithmic policies implemented by the public administration.

A public body that aims to adopt a certain algorithmic policy must use the impact assessment process both to provide information and justifications to society

¹⁶⁶Reisman et al. (2018), pp. 18–19.

¹⁶⁷Article 48 of the Draft Bill 2338/2023.

¹⁶⁸Article 48, IX of the Draft Bill 2338/2023.

¹⁶⁹Reisman et al. (2018), p. 10.

regarding the system it aims to deploy. Through AIAs, public bodies are invited to submit themselves to public scrutiny, having the opportunity to transform or correct their practices and adopt policies that are more responsive to the needs and rights of the population.

If AIA is to be an effective instrument for democratic and social control of algorithmic policies implemented by states, it has to embrace participatory processes. AIAs must include mechanisms for public consultation and contestation, so that researchers, citizens, and civil society organizations can dispute the definition of impacts and risks, contest specific information presented in the initial assessment report, or even disagree with the implementation of the system as a whole.

Although participatory processes come with trade-offs, it's important to consider that if appropriately designed, their benefits outweigh their drawbacks. While public consultations make the process more complex and time-consuming, they also increase awareness of the social context in which the AI system will be implemented. Additionally, they improve the chances of identifying risks beforehand and reduce the likelihood of discriminatory biases, thus better protecting the fundamental rights of affected individuals and communities.

However, participation is rarely observed in practice. In general, the role of society in AIA processes is still underexplored. There are few economic and institutional incentives for accountability practices to incorporate actors beyond the internal scope of the companies that manufacture the algorithms and the public agencies that hire or develop them.

As revealed by the research carried out by the Ada Lovelace Institute on some jurisdictions, such as in New Zealand and Canada, the impacted communities had very little influence on the outcome of the impact assessments,¹⁷⁰ especially concerning how the potential harms of the systems were identified and addressed. As discussed by the report, “owing to their limited transparency and public participation, most AIAs currently implemented fall substantially short of models for AIAs proposed by experts”.¹⁷¹

Some AIA processes already deployed by the public sector, as is the case of Canada, are mainly designed to determine the algorithmic system's level of risk, being limited to an internal analysis of the bodies that aim to implement them, without a public debate that enables the weighting-up costs and benefits of the measure.¹⁷² This situation raises two main concerns: the lack of social participation in AIA processes, and the reduction of AIA to a mere instrument for risk classification. All indications are that the current practice and institutionalization of AIAs fall short of many relevant functions pointed out by this work, such as mitigating damage, weighing up costs and benefits, and providing accountability and social control.

Concerning the European AI Act, although it mandates FRIAs as a compulsory process for high-risk systems, it fails to provide any degree of public transparency

¹⁷⁰ Ada Lovelace Institute et al. (2021), p. 21.

¹⁷¹ Ada Lovelace Institute et al. (2021), p. 23.

¹⁷² Ada Lovelace Institute et al. (2021), p. 21.

regarding the results of the assessments. Furthermore, the approved version of the Act does not address the necessity for social participation in AIA processes.

The lack of mechanisms for transparency and social participation in AIAs in significant regulatory initiatives, such as the AI Act, is a warning that social control over AIAs should receive more academic and regulatory attention. Therefore, there is a growing need to invest in research aimed at developing meaningful and efficient participatory methods for AIA.

The Brazilian proposal is on the right track by demanding the involvement of various social groups in AIA processes, particularly those vulnerable groups that could be impacted by AI systems. Requiring deployers to disclose AIA results through easily accessible communication channels is also a significant step towards social oversight of AI systems, which is also included in the Brazilian proposal. However, it's important to note that the bill is currently under review by Congress and may still be modified.

As AIA is still a relatively new instrument, it is expected that AIA regimes in the public sector will develop over time, based on a combination of legislation, regulations, case law, public contributions, and academic studies.¹⁷³

As proposed in this contribution, some criteria are important for the institutionalization of AIA, such as the establishment of *binding rules*, a *supervisory authority*, and its *scope of application*. Its *implementation process* should also be regulated. Impact assessment processes ideally should include not only the preparation of an internal self-assessment by suppliers and system operators but also stages for disclosing the assessment report in easily accessible vehicles, for public consultation and contestation.

As Moss et al state: “Attending to how AIA processes are assembled is imperative because they may be the means through which a broad cross-section of society can exert influence over how algorithmic systems affect everyday life”.¹⁷⁴ How AIA processes are regulated is of fundamental importance to determine if and how the purposes of accountability and social control on governments’ use of AI will be achieved in the future.

References

- Access Now (2024) Radiografía Normativa: ¿Dónde, Qué Y Cómo se está regulando la Inteligencia Artificial en América Latina?. Access Now. <https://www.accessnow.org/wp-content/uploads/2024/02/LAC-Reporte-regional-de-politicas-de-regulacion-a-la-IA.pdf>
- Ada Lovelace Institute, AI Now Institute, Open Government Partnership (2021) Algorithmic Accountability for the Public Sector. Open Government Partnership. <https://www.opengov-partnership.org/documents/algorithmic-accountability-public-sector/>
- Agencia Senado (2024) Votação de regulamentação de IA é adiada pela terceira vez. <https://www12.senado.leg.br/noticias/materias/2024/07/09/votacao-de-regulamentacao-de-ia-e->

¹⁷³ Moss et al. (2021), p. 14.

¹⁷⁴ Moss et al. (2021), p. 5.

- adiada-pela-terceira-vez#:~:text=%C3%89%20a%20terceira%20vez%20que,buscam%20proteger%20as%20suas%20atividades
- Algorithm Watch (2020) How Dutch activists got an invasive fraud detection algorithm banned. <https://algorithmwatch.org/en/syri-netherlands-algorithm/>
- Amnesty International (2021, October 25). Xenophobic Machines: Discrimination through unregulated use of algorithms in the Dutch childcare benefits scandal. Amnesty International. <https://www.amnesty.org/en/documents/eur35/4686/2021/en/>
- Amrute S, Ranjit S, Rigoberto L (2022) A primer on AI in/from the majority world: an empirical site and a standpoint. Data & Society Research Institute. https://datasociety.net/wp-content/uploads/2022/09/09142022_AIMW_Primer_eng_final.pdf
- Barcellos A (2015) Acesso à informação: os princípios da Lei nº 12.527/2011. *Revista Quaestio Iuris* 8(3). <https://doi.org/10.12957/rqi.2015.18818>
- Bioni B, Luciano M (2020) The precautionary principle in the regulation of artificial intelligence: would data protection laws be its gateway? In: Frazão A, Mullholand C (eds) *Artificial intelligence and law: ethics, regulation, and responsibility*. *Revista dos Tribunais*
- Bovens M (2007) Analyzing and assessing accountability: a conceptual framework. *Eur Law J* 13(4):447–468. <https://doi.org/10.1111/j.1468-0386.2007.00378.x>
- Brandusescu E, Reia J (eds) (2022) *Artificial Intelligence in the city: building civic engagement and public trust*. Centre for Interdisciplinary Research on Montreal, McGill University. https://www.mcgill.ca/centre-montreal/files/centre-montreal/artificial_intelligence_in_the_city.pdf
- Buolamwini J, Gebru T (2018) Gender shades: intersectional accuracy disparities in commercial gender classification. In: *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, pp 77–91. <https://proceedings.mlr.press/v81/buolamwini18a.html>
- Burrell J (2016) How the machine ‘thinks’: understanding opacity in machine learning algorithms. *Big Data Soc* 3(1). <https://doi.org/10.1177/2053951715622512>
- Buschmann J (2021) *Sistema Preditivo do Crime Urbano: Produção algorítmica de zonas de vigilância e controle na cidade*. *Derechos Digitales*. <https://www.derechosdigitales.org/publicaciones/sistema-predictivo-do-crime-urbano-producao-algoritmica-de-zonas-de-vigilancia-e-controle-na-cidade/>
- Costa R, Kremer B (2022) Inteligência artificial e discriminação: desafios e perspectivas para a proteção de grupos vulneráveis frente às tecnologias de reconhecimento facial. *Revista Direitos Fundamentais e Justiça* 16(1). <https://doi.org/10.30899/dfj.v16i1.1316>
- Danaher J. (2016) The threat of algocracy: reality, resistance and accommodation. *Philos Technol* 29(3):245–268. <https://doi.org/10.1007/s13347-015-0211-1>
- Data & Society & European Center for Not-for-Profit Law. *Mandating Human Rights Impacts Assessments in the AI Act*. <https://ecn1.org/sites/default/files/2021-11/HRIA%20paper%20ECNL%20and%20Data%20Society.pdf>
- Eubanks V (2018) *Automating Inequality: How High-tech Tools Profile, Police, and Punish the Poor*. St. Martin’s Press.
- European Commission (2015) *Better regulation guidelines*. Eur-Lex. <https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52015SC0111&rid=1>
- European Commission (2019) *Ethics guidelines for trustworthy AI | Shaping Europe’s digital future*. European Commission. <https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai>
- Fernández J (2022) Artificial intelligence in government: risks and challenges of algorithmic governance in the administrative state. *Indiana J Glob Leg Stud* 30(1):65–95. <https://doi.org/10.2979/gls.2023.a886163>
- Floridi L (2021) Artificial intelligence as a public service: learning from Amsterdam and Helsinki. *Philos Technol* 33(4):541–546. <https://doi.org/10.1007/s13347-020-00434-3>
- Frajhof I (2022) *Right to explanation and protection of personal data in decisions by artificial intelligence algorithms*. Doctoral dissertation, Pontifical Catholic University of Rio de Janeiro
- Gerards J, Schäfer M, Vankan A, Muis I (2022) *Impact assessment fundamental rights and algorithms*. Ministry of the Interior and Kingdom Relations. <https://www.government.nl/documents/reports/2021/07/31/impact-assessment-fundamental-rights-and-algorithms>

- Ghaffary (2019) San Francisco's facial recognition technology ban, explained. Vox. <https://www.vox.com/recode/2019/5/14/18623897/san-francisco-facial-recognition-ban-explained>
- Haataja M et al (2020) Public AI Registers Realizing AI transparency and civic participation in government use of AI. <https://ai.hel.fi/wp-content/uploads/White-Paper.pdf>
- Jewell M (2018) Contesting the decision: living in (and living with) the smart city. *Int Rev Law Comput Technol* 32(2–3):210–229. <https://doi.org/10.1080/13600869.2018.1457000>
- Kaminski M (2019) Binary governance: lessons from the GDPR's approach to algorithmic accountability. *South Calif Law Rev* 92(6) <https://scholar.law.colorado.edu/faculty-articles/1265/>
- Karlin M (2018) The Government of Canada's Algorithmic Impact Assessment: Take Two. Medium. <https://medium.com/@supergovernance/the-government-of-canadas-algorithmic-impact-assessment-take-two-8a22a87acf6f>
- Keller P (2019) Participatory accountability at the dawn of artificial intelligence. King's College London Law School Research Paper 31. <https://doi.org/10.2139/ssrn.3448315>
- Krasmann S (2020) The logic of the surface: on the epistemology of algorithms in times of big data. *Inform Commun Soc* 23(14):2096–2109. <https://www.tandfonline.com/doi/abs/10.1080/01369118X.2020.1726986>
- Lemos A, Buarque G, Soares I, Mulin V, Chiavone T (2023) Avaliação de Impacto Algorítmico para a proteção dos direitos fundamentais. Laboratório de Políticas Públicas e Internet. <https://lapin.org.br/wp-content/uploads/2023/04/RelatorioAIA.pdf>
- Lessig L (1999) Code and other laws of cyberspace. Basic Books, New York City
- Lynch C (2020) Contesting digital futures: urban politics, alternative economies, and the movement for technological sovereignty in Barcelona. *Antipode* 52(3):660–680. <https://doi.org/10.1111/anti.12522>
- Mantelero A (2024) The fundamental rights impact assessment (FRIA) in the AI act: roots, legal obligations and key elements for a model template. *Comput Law Secur Rev* 54:106020. <https://doi.org/10.1016/j.clsr.2024.106020>
- Meneguín F, Silva R (2017) Introdução. In: Meneguín F, Silva R (eds) Avaliação de Impacto Legislativo: Cenários e perspectivas para sua aplicação. Senado Federal, Coordenação de Edições Técnicas, pp 14–23. <https://www2.senado.leg.br/bdsf/handle/id/535244>
- Mökander J, Axente M, Casolari F, Floridi L (2021) Conformity assessments and post-market monitoring: a guide to the role of auditing in the proposed European AI Regulation. *Minds Mach* 32(2):241–268. <https://doi.org/10.1007/s11023-021-09577-4>
- Morozov E, Bria F (2018) Rethinking the Smart city: democratizing urban technology. Rosa Luxemburg Stiftung, Berlin
- Moss E, Watkins EA, Singh R, Elish MC, Metcalf J (2021) Assembling accountability: algorithmic impact assessment for the public interest. *SSRN Electronic Journal*. <https://doi.org/10.2139/ssrn.3877437>
- Mulholland C, Frajhof I, Guedes P, Iunes J (2023) Avaliação de impacto algorítmico: o que é e como está regulada no PL 2.338/23 do Brasil. Migalhas de Responsabilidade Civil. <https://www.migalhas.com.br/coluna/migalhas-de-responsabilidade-civil/395547/avaliacao-de-impacto-algoritmico>
- Nemitz, P. (2018). Constitutional democracy and technology in the age of artificial intelligence. *Royal Society Philosophical Transactions A*. <https://doi.org/10.2139/ssrn.3234336>
- Newman J, Mintrom M (2023) Mapping the discourse on evidence-based policy, artificial intelligence, and the ethical practice of policy analysis. *J Eur Public Policy* 30(9):1839–1859. <https://doi.org/10.1080/13501763.2023.2193223>
- O'Donnell G (2004) The quality of democracy: why the rule of law matters. *J Democr* 15(4):32–46. <https://doi.org/10.1353/jod.2004.0076>
- O'Neil C (2016) Weapons of math destruction: how big data increases inequality and threatens democracy. Crown publishing group.
- Oxford Insights (2023) Government AI readiness index. Oxford Insights. <https://oxfordinsights.com/ai-readiness/ai-readiness-index/>

- Pasquale F (2015) *The black box society: the secret algorithms that control money and information*. Harvard University Press
- Reisman D, Schultz J, Krawford K, Whittaker M (2018) *Algorithmic impact assessments: a practical framework for public agency accountability*. AI Now Institute. <https://ainowinstitute.org/publication/algorithmic-impact-assessments-report-2>
- Rodrigues G (2019) *Reconhecimento Facial na Segurança Pública: Controvérsias, riscos e regulamentação*. Instituto de Referência em Internet e Sociedade. <https://irisbh.com.br/reconhecimento-facial-na-seguranca-publica-controversias-riscos-e-regulamentacao/>
- Selbst A (2021) An institutional view of algorithmic impact assessments. *Harv J Law Technol* 117. <https://ssrn.com/abstract=3867634>
- Sheard N, Schwartz A (2022) *The movement to ban government use of face recognition*. Electronic Frontier Foundation. <https://www.eff.org/deeplinks/2022/05/movement-ban-government-use-face-recognition>
- Smuha NA (2021) Beyond the individual: Governing AI's societal harm. *Internet Policy Rev* 10(3). <https://doi.org/10.14763/2021.3.1574>
- Smuha N (2022) *The algorithmic leviathan and the rule of law: an examination of how algorithmic regulation in the public sector can lead to 'algorithmic rule by law', and a critical evaluation of EU legal safeguards*. [Doctoral dissertation, KU Leuven Faculty of Law]
- van Bekkum M, Borgesius F (2021) Digital welfare fraud detection and the Dutch SyRI judgment. *Eur J Soc Secur* 23(4):323–340. <https://doi.org/10.1177/13882627211031257>
- Wieringa M (2020) What to account for when accounting for algorithms: a systematic literature review on algorithmic accountability. In: *FAT* '20: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. <https://dspace.library.uu.nl/handle/1874/414522>
- Worldbank (2021) *Data for better lives*. Washington, DC. <https://www.worldbank.org/en/publication/wdr2021>
- Zouridis S, van Eck M, Bovens M (2019) Automated discretion. In: Evans T, Hupe P (eds) *Palgrave handbook on discretion: the quest for controlled freedom*. Springer International Publishing, pp 313–329. <https://ssrn.com/abstract=3453068>

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.



As Much as Faith: A Speculation on Quantum Computing with Fiduciary Law in Public Governance



Robert Herian

Abstract Quantum computing is arguably the most sublime of ‘emerging’ technologies. By harnessing the laws of quantum mechanics, quantum computing can solve problems too complex for classical computers and do so with much greater speed and efficiency. Yet, despite the opportunities, quantum computing heralds another stage of human technological innovation that threatens to evade human know-how, understanding and, therefore, control. Quantum computing in the service, management, and governance of societal needs and desires must, therefore, embrace “known unknowns” and engender leaps of faith—*quantum fide* (as much as faith). At stake are corresponding needs and desires to identify and safeguard core public values and legal norms when designing and implementing quantum computing as part of a broader socio-technical infrastructure. In speculating how best to define and address the challenges and opportunities quantum computing offers public governance, this chapter considers fiduciary law as a body of theory and practice broadly directed towards acting on behalf of others and helping relieve the burden of know-how of increasingly complex technologies. Fiduciary law could provide an answer to how quantum computing in public governance might proceed from a promissory idea to development and deployment in a legally, ethically, and socially acceptable way.

R. Herian (✉)
University of Exeter Law School, Exeter, UK
e-mail: r.herian@exeter.ac.uk

1 Introduction

Despite clear opportunities, quantum technologies (QT) and quantum computers (QC), which ‘harnesses some of the almost-mystical phenomena of quantum mechanics to deliver huge leaps forward in processing power’,¹ herald yet another stage in human technological progress that threatens to evade understanding and control at many levels of operation. This matters because it must be possible to employ a robust, flexible, and enforceable public governance framework, a set of rules able to manage the interests of stakeholders, check opportunism, and ensure QC implementations are in the public interest and confer public benefit. This applies to the public sector but equally concerns private sector actors seeking to influence and profit from public sector contracts. Speculating on how to define and address the challenges and opportunities QC offers to governance networks, public services, and administrative decision-making procedures required to manage them, this contribution looks to existing principles and doctrines drawn from the common law of England and Wales.

Specifically, fiduciary governance (as a variant or species of fiduciary law administered in equity),² which I argue supports a general need and desire for ‘developing dynamic governance models that respond, adjust and update according to the changing status and impact of technology.’³ Echoing the law,⁴ fiduciary governance is defined for present purposes, as a collection of rights and corresponding obligations whereby an agent (the fiduciary) acts with fidelity and without self-interest on behalf of another (the principal) to meet and master the principal’s needs and thus reduce or relieve burden.⁵ Failure to meet the needs of the principal may result in a breach from which flows a suite of available remedies for the claimant, including compensation and restitution.

The deep-rooted, pre-Christian, Latin etymology of the fiduciary (*fiducia*: embodying trust and confidence) in conjunction with notions of fidelity (*fides*: embodying faith, honesty, etc.) and theological characteristics of the notion of burden (as well as equity, although that subject will not be addressed at length here),⁶

¹ See Giles (2019).

² Fiduciary law, as a body of legal doctrine, developed over the course of several centuries in the courts of equity (Chancery) of England and Wales. As a body of law, equity itself has deep roots in ancient philosophy, most notably Aristotle, filtered through Christian doctrine before settling as a secular law and mode of contemporary justice (see, for example, Herian (2021)).

³ Perrier (2022), p. 17.

⁴ Fiduciary governance is said to differ from law because the focus is not on the interests or property of identifiable individuals, but rather particular purposes (Miller and Gold 2015)).

⁵ This differs from the regulation of technology using a generalized framework of rights, interests, and duties because of the specific alignment of fiduciary obligations with the notion of burden. For a more generalized framework, albeit one relating to QC, see: Perrier (2022).

⁶ See, for example, Herian (2022).

are all key to the position taken in this contribution.⁷ My account of meeting and mastering the principal's needs and burden is predicated on an entirely secular (profane) property of fiduciary governance, which can be realised in the governance and regulation of QC deployed by a range of stakeholders, including public agencies and private actors. Further, fiduciary governance can, indeed ought in time, to be extended to non-human, machine-to-machine actors, or agents.⁸

Risks from QC and its applications stemming from “known unknowns” as well as “unknown unknowns” require engineers, designers, and implementors (collectively “stakeholders”) to have significant skills in explainability to mitigate potential harms. Like artificial intelligence (AI) but with the possibility for even less human operational certainty, QC marshals the forces of nature in such a way as to render the processes and products of computation enigmatic if not entirely alien. As the eminent physicist Richard Feynman said, ‘Things on a very small scale [the atomic scale of “quantum mechanics”] behave like nothing that you have any direct experience of. They do not behave like waves, they do not behave like particles, they do not behave like clouds, or billiard balls, or weights or springs, or like anything that you have ever seen.’⁹

Turning to quantum mechanics in the service, management, and governance of societal needs and desires offers opportunities and challenges regarding both broader forms of QC and narrower and more specific forms of quantum simulation.¹⁰ In either case, applied outside of controlled laboratory environments, various unknowns concerning the technology must be expected. Therefore, a need for citizens to have belief or faith in the stakeholders of quantum hardware and systems—*quantum fide* (as much as faith)—arises, especially where knowledge gaps not only persist but widen. It matters because, much like the unfolding story of AI, QC will not only supplement but likely supersede humanity's capacity to understand and control online (virtual), offline (real world), and hybrid environments built on or around QT and QC. As Valentin Jeutner claims, ‘there is no doubt that the remarkable potential of quantum computing will, if unchecked, impact social

⁷ See, for example, The Hon Justice James Edelman (2014). References to the theological or sacred is arguably important because it can account for degrees of evangelism that often accompany innovation and entrepreneurialism in contemporary Western capitalist contexts. That is not to say, however, that the aim of this analysis or its conclusions is to prove or show an explicit continuation of organized religious thought in or through public governance and its application of technologies.

⁸ In other words, it is necessary to ask the question beyond the scope of the present chapter of whether a quantum algorithm coded with fiduciary governance principles and running without human intervention or oversight is desirable or possible as the technology advances into wider, everyday contexts, and what this might look like in practice.

⁹ Feynman (1998), p. 116.

¹⁰ ‘In the near term, quantum simulation is the most interesting form of quantum computing. Quantum simulators could be understood as quantum computers designed for solving one type of problems. These quantum systems are engineered to run a specific algorithm. Hence, less operations are needed than for a machine that is aimed at solving more general questions.’ (De Jong (2022), p. 6).

power-relations in a remarkable manner.¹¹ Coming to terms with and accounting for the novelties of QC is therefore key to ensuring a reasonable and beneficial balance between humanity and the technology. Through fiduciary governance, as a set of values and principles encompassing interrelational and transactional notions of loyalty, prudence, care, and trust, this contribution speculates on how QC in public governance might proceed from a promissory idea to development and deployment in a legally, ethically, and socially acceptable way.

To begin, I want to introduce in more detail why burden is key to understanding deployment of new innovations such as QC in public (citizen) as well as private (consumer) settings. Following this I will turn to an overview of quantum computing, including how these machines operate in contrast to classical computers and what is at stake from their rise to prominence. Next this contribution will consider key features of fiduciary governance before merging these ideas with characteristics of quantum computing to discuss how the two could work together in the service of governance's aims and objectives in public as well as private sectors.

2 The Burden of Knowing

Knowledge is often considered an asset, a source of power, perhaps even still a path to enlightenment.¹² Yet, beneath the surface of its many benefits lies, I suggest, a less-discussed aspect: knowing, wanting, or needing to know, and feeling unable to keep up with the pace of innovation creates a burden. This burden can take many forms, from the emotional weight of difficult truths to the responsibility that comes with awareness and effects a wide range of individuals and communities, notably those with vulnerabilities.¹³ Burden is a double-edged sword that can shape lives in profound ways. In our information age, we are bombarded with a constant stream of data, facts, and opinions. The more we know, the more we are expected to process and make sense of this information. As we accumulate knowledge, we may find ourselves carrying the weight of facts and statistics, struggling to filter out what is relevant and meaningful. The more we know, the more we must confront the vastness of what we don't know, and it is here that we find the likes of belief and faith, whether sacred or profane, filling gaps in knowledge.¹⁴

QC in public governance comes with corresponding needs and desires to identify and safeguard core public values and legal norms when designing and

¹¹ Jeutner (2021), p. 53.

¹² See, for example, Foucault (2002), pp. 111–133.

¹³ Vulnerability is not a theme dealt with expressly here. However, it is important to recognize vulnerability and theories that describe how it impacts and shapes society as key to understanding burden more broadly and would be required to explore a theory of burden more fully. To echo Martha Fineman's (2008) work, therefore, we may suggest the need to better understand the *burdened subject* as an alternative to the normative liberal subject.

¹⁴ See for example, Knapp and Michaels (1982), pp. 736–742.

implementing broader socio-technical infrastructure. As the World Economic Forum (WEF) identifies in its initial governance principles for QC, ‘With quantum computing introducing a new computing paradigm, existing knowledge about the opportunities and risks of a new technology as well as the necessary understanding to adjudicate between them will, in many cases, be insufficient or at least call for a reassessment of how such opportunities and risks are managed.’¹⁵ Explainability or a more generalised concept of “education” is often regarded as important, even a panacea, for regulators and policymakers keen to mitigate widening public knowledge gaps in the wake of innovation. This includes what is often referred to as “digital” or “online media” literacy, seen by the UK government in recent years as necessary to ensure full participation in society as engaged and empowered citizens.¹⁶ Hence, it is now common practice to point individuals and communities towards different sources of learning and greater levels of (online) information. This reflects the fact that the onus is almost always on those citizen’s subject to new and emerging technologies in public (and private) sectors to understand the impacts on themselves for themselves or risk, to echo the UK government position outlined above, partial or less than full participation in society. The bargain, therefore, is relatively straightforward: if a citizen does not engage, they should not expect to be empowered.

Knowledge and understanding are seen as crucial for informed decisions to be made and, especially, for consent to be reasonably given when it comes to use, exposure, or engagement with technologies and applications, and a vast academic literature has grown around the topic.¹⁷ Central to policy is a notion of “empowerment” through “literacy”, although this is primarily regarding consumer contexts of use, exposure, or engagement rather than in-depth know-how of the technologies as such.¹⁸ The conclusion reached is that a failure for government to suitably raise the technological literacy levels of citizens (albeit as consumers) risks at best confusion regarding technologies (what they do and, importantly, why), at worst loss (personal, financial, etc.) and harm.

How feasible is the task of education (“literacy”, “empowerment”, etc.) given the complexities of QC and a concern that ‘gives rise to distinct governance considerations that may not arise in other fields.’¹⁹ Perhaps we should aim for something vaguer or less concrete than education in this case, such as “demystification”?²⁰ Where public or government agencies and departments plan to use QC to meet and master the wicked and messy realities of daily governance, explainability, as education or demystification, is fraught with problems and unlikely to achieve its goals.

¹⁵ WEF (2022a), p. 4.

¹⁶ Department for Digital, Culture, Media & Sport (2022), p. 23.

¹⁷ Tinmaz et al. (2022).

¹⁸ See, for example, the Information Commissioner’s Office (ICO) in the UK, whose role is to provide information to data subjects regarding their rights over personal data in technological settings (e.g., AI) that require such data to function. <https://ico.org.uk/about-the-ico/what-we-do/>.

¹⁹ WEF (2022a), p. 5.

²⁰ See De Jong (2022).

Explaining highly complex technologies deployed to regulate heterogeneous, contingent environments, which includes an increasingly broad range of software, hardware, and infrastructure, all with the potential to radically impact lives and livelihoods, is, I argue, not feasible at all. A key reason for this is, as Alexander Galloway has described it, that ‘What was once a question of consideration and sense is now a question of logic and physics’.²¹ Applied to the example of the rather banal notion of the “back-office”, which describes systemically embedded and often invisible computer operations in public or private institutions and bureaucracies, we find a creeping tendency towards obscurity and obfuscation no-longer concerned with consideration and sense. Those who do not understand logic and physics, who may prefer consideration and sense, are at best excluded and worst left unrepresented by shifts in innovation and the emergence of new systems and ways of working and organizing society.

It is a longstanding part of the human condition to want to meet an infeasible solution to a problem by imagining and constructing the expertise and specialism to solve it. But this may not be feasible for most individuals or communities, especially those facing challenges economically and due to time poverty, the latter of which is central to the increase of, among other things, psychological burdens.²² This point may seem trite, but in the wake of widespread criticism of “experts” and technocracy more broadly, following perceived and, often, actual failures by governments of all political type to foster or maintain healthy and equitable societies, the need to assuage vulnerabilities and relieve burdens in society is arguably greater than it has ever been.²³ It is a burden on individuals, communities, and government to ensure the requisite education to close knowledge gaps in the hope of improving know-how concerning many contemporary soft- and hardware technologies, when a solution could (or should) be to ensure oversight of technologies with public impact comply with the priority for public benefit.

Burden in this context must not be mistaken, therefore, with a general, or even specified unwillingness, neglect, or laziness for individuals or communities to learn how to use, repair, or manage technology. Rather, burden describes a limit to the overall capacity for know-how, often expressed against a measure of time, but not always or exclusively. Nonetheless, we may account for a “burden of knowing” within the ambit of wider (personal) administrative burdens that are increasingly placed on individuals (citizens) and communities in advanced capitalist economies and jurisdictions such as the UK and US, where that burden disproportionately falls

²¹ Galloway (2004), p. 7.

²² See Giurge et al. (2020).

²³ The aim here is not a full-throated defense of technocracy, especially as it can do and does conflict with democracy, to the detriment of the many. Equally, however, to dismiss the importance of expertise under the assumption that individuals can or should navigate the world under, what I argue is, an illusory aegis of “(self-)empowerment” is also unhelpful. An answer may lie on a spectrum of epistemic democratic theories and practices that neither wholly reject elements of technocracy nor assume fully participatory “empowered” modes of democracy. For discussion on some of these points, see Cole (2022).

on those in lower income brackets.²⁴ I argue that the limit that surrounds and defines a burden is not fixed but is elastic in the context of personal administration.²⁵ Burden, therefore, speaks to a boundary that flexes as the burden pushes an individual or community to new, different, and often uncomfortable positions and conclusions. In this sense, burden describes an inflating and contracting in space and time relative to use, exposure, and engagement with innovations and technologies.

Whilst incorporating certain types of expertise or specialism, burden also and perhaps more straightforwardly describes the intermediation and intervention of another who is willing and able to share or assume the burden in part or in full; an intervention that both recognises the limit the burden inscribes and acts on it. If the task of educating or attaining reasonable know-how on QC and its implementation and use in common environments and settings is a burden, therefore, how should this burden and the limit it entails be measured and addressed to ensure, among other things, the development and deployment of QC in a legally, ethically, and socially acceptable way? The means suggested here is fiduciary governance. But before turning to look at fiduciary governance in more detail, it is important to explore some basics of QC and consider what is at stake from the operational capabilities of this nascent technology.

3 Quantum Computing (the Basics)

For present purposes we are interested in those properties of QC that may, in time, work with classical computing or independent of it to support governance aims and objectives, rather than a discussion of the details or the *long durée* of quantum mechanics. On quantum mechanics it is enough to say that it is ‘ubiquitous in modern technology’ such as ‘semiconductors, transistors, microchips, lasers, and computer memory’;²⁶ notwithstanding the variety of theoretical excursions the field has unleashed since formalization in the 1920s²⁷—including the infamous “many worlds” theory. Quantum mechanics is viewed by physicists such as Carroll as ‘the deepest, most comprehensive view of reality we have’, meaning that, ‘quantum mechanics isn’t just an approximation of the truth; it is the truth.’²⁸

²⁴Giurge et al. (2020), p. 999.

²⁵In terms of the role played by technology, the notion of “personal administration” as I am referring to it here overlaps significantly with the notion of “governmentality” as I have used that term elsewhere in relation to blockchain technologies (Herian (2019), p. 29).

²⁶Carroll (2019), pp. 1–2.

²⁷‘By 1928 or so, many of the mathematical formulas and rules of quantum mechanics had been put in place and, ever since, it has been used to make *the* most precise and successful numerical predictions in the history of science.’ (Greene (2000), p. 87).

²⁸Carroll (2019), p. 2.

3.1 *Manipulating Quantum Bits*

At the heart of QC are quantum bits or *qubits*, the quantum analogue of classical bits (0 and 1). Qubits can represent both 0 and 1 simultaneously, thanks to a phenomenon known as superposition. This allows quantum computers to perform many calculations in parallel, making them inherently more powerful for certain tasks. In classical computing, bits are either 0 or 1, but in QC, qubits can exist in a superposition of both states simultaneously. As such, quantum computers ‘manipulate qubits in the same way that ordinary computers manipulate classical bits.’²⁹ This property enables a QC to explore multiple solutions to a problem simultaneously and why applications of the technology in a wide variety of social contexts are interesting and exciting propositions.

A major property of qubits is entanglement. “Quantum entanglement”, as Preskill describes it, ‘is the principle that even if we know everything about the whole system, you don’t necessarily know everything about its parts. Classical systems are not like that. In the classical world, if I know everything that can be known about the full system, I know everything about each of its parts as well.’³⁰ When qubits become entangled, the state of one qubit instantly influences the state of another, regardless of the physical distance between them, meaning that when two systems are entangled there ‘exists a special connection between them.’³¹ This property allows for the potential for secure communication and the development of quantum cryptography. As Kop et al maintain, ‘Entanglement is directly harnessed in QC along with superposition to achieve quantum parallelism resulting in quantum algorithms with exponential speed-up over classical computations. It is also harnessed in quantum communications by taking advantage of unique correlations exhibited by entangled qubits and quantum cryptography, particularly quantum key distribution (QKD)’.³²

Secure communication in the context of QC is a pressing and evolving challenge. As QC technology advances, it has the potential to both threaten and enhance the security of our communication systems, including the potential to disrupt classical cryptographic methods that rely on the difficulty of solving complex mathematical problems. Traditional encryption, such as Rivest-Shamir-Adleman (RSA) and Elliptic-curve cryptography (ECC), could be vulnerable to quantum attacks, as quantum computers may efficiently factor large numbers or solve discrete logarithm problems. As a result, secure communication in the age of QC demands new cryptographic techniques that are resistant to quantum attacks.

One area in which the impacts of QC on cryptography are already being taken seriously and are being addressed, is blockchain and cryptocurrency, where whole decentralized networks, both private and public, could be undermined. As Davis and Kim explain:

²⁹ Carroll (2019), pp. 83–84.

³⁰ Preskill (2021), p. 11.

³¹ Kop et al. (2023), p. 5.

³² Kop et al. (2023), p. 5.

cryptocurrency miners using quantum computers may be able to mine cryptocurrency much faster than other miners. This could threaten the decentralization of many mining-based blockchain protocols. Quantum computers could also decrypt the private key from a public key, enabling bad actors to control, and ultimately steal, others' cryptocurrencies. Quantum computers thus could pose a significant threat to cryptocurrencies and blockchain technology. In response, some developers are already working to future-proof their blockchain protocol by exploring ways to transition the cryptography currently securing the protocol to quantum-resistant cryptography.³³

Mentioned above by Kop et al., quantum cryptography, also known as quantum key distribution (QKD), is one of the most promising approaches to achieving secure communication in a post-QC world. QKD can establish a secure key between two parties that cannot be intercepted without detection, and this is something blockchain developers are utilising to meet the challenges QC pose to blockchains. QKD protocols such as the BBM92 involve transmission of qubits, usually in the form of photons, between the sender and receiver.³⁴ Any attempt to intercept these qubits will be immediately detected due to the disturbance caused by measurement, ensuring the security of the key exchange. Furthermore, the no-cloning theorem in quantum mechanics states that an arbitrary unknown quantum state cannot be copied exactly.³⁵ This property adds an additional layer of security, as it prevents an eavesdropper from creating an undetectable copy of the quantum states being transmitted. Despite the inherent security of QKD, it is not without its challenges. Practical QKD systems require specialized hardware, such as single-photon detectors and quantum key generators. Additionally, QKD systems, like QC more broadly in its present forms, are susceptible to various practical issues like channel loss and noise, which can limit their range and efficiency.

3.2 *Algorithms and Applications*

As we have seen, QC has the potential to outperform classical computers in specific domains, thereby engendering different levels of advantage and supremacy.³⁶ Where narrowly defined tasks are assigned to QC, referred to earlier as particular “simulations”, alone or in conjunction with classical computing, advantage or supremacy will represent an overall intensification in capabilities and outputs. This is perhaps most notable with quantum algorithms, as well as in a variety of applications they allow for, especially those in which QC works together with, for instance, machine learning and artificial intelligence. Key examples are Shor's algorithm, which can factor large numbers exponentially faster than classical algorithms, with significant implications for breaking current encryption standards. This poses both a security

³³ See Davis and Kim (2023).

³⁴ See Bennett et al. (1992).

³⁵ See Yao et al. (2014).

³⁶ Atik (2022), Perrier (2022).

risk and an opportunity for the development of post-quantum cryptography, as discussed above. Further, Grover's algorithm speeds up the process of searching an unsorted database, offering a quadratic speedup compared to classical algorithms. These polynomial resources can have a substantial impact on optimization problems, which translate into significant advantages in information retrieval and processing for sectors that are heavily reliant on detailed information, such as legal practice or governance.³⁷

Quantum bits are highly susceptible to errors from various sources, including environmental factors and imprecise gate operations.³⁸ Developing robust error correction codes is crucial for making quantum computers reliable. Further, quantum states are fragile and tend to lose coherence quickly. Maintaining the coherence of qubits over extended periods is essential for the practical implementation of quantum algorithms. As Preskill outlines, 'physicists were deeply familiar with the debilitating effects of decoherence, which under typical conditions prevent complex many-particle quantum systems from exhibiting quantum behaviour, and viewed quantum computing (in the derisive words of Haroche and Raimond) as the computer scientist's dream [but] the experimenter's nightmare.'³⁹ 'Remarkably', Preskill continues,

it was again Shor who led the next pivotal advances the discovery of quantum error-correcting codes and of fault-tolerant methods for executing a quantum computation reliably using noisy hardware. By the end of 1996, it was understood at least in principle that quantum computing could be scaled up to large devices that solve very hard problems, assuming that errors afflicting the hardware are not too common or too strongly correlated.⁴⁰

It is via the likes of Shor's and Grover's algorithms, therefore, that we can further apprehend (if not fully comprehend) the enigmatic or entirely alien capabilities that quantum computers present to us. As Atik maintains regarding QC applied to the law, it 'will expand possibilities of modelling law, will introduce new regulatory forms, and will lead to more powerful tools for the legal profession. *Many of these changes will take place "under the hood" and will not demand a thorough understanding of quantum theory, but the practitioner will experience a different feel from the technology. Legal information retrieval will bring back different—and more useful—results.*' [emphasis added]⁴¹

³⁷ Atik and Jeutner (2021), p. 304.

³⁸ Quantum computation is carried out through quantum gates, analogous to classical logic gates, which manipulate qubits, allowing for various operations and computations. Common quantum gates include the Hadamard gate, CNOT gate, and the phase gate. The Hadamard gate puts qubits into superposition, allowing them to explore both 0 and 1 states simultaneously, whilst the CNOT gate (controlled-NOT gate) is essential for entangling qubits, making it a fundamental building block for many quantum algorithms. Finally, the phase gate introduces phase shifts to qubits, influencing their amplitudes and enabling interference effects in quantum algorithms.

³⁹ Preskill (2021), p. 5.

⁴⁰ Preskill (2021), p. 5.

⁴¹ Atik (2022), p. 16.

The development and adoption of QC have far-reaching implications for various sectors and ‘is seen as a strategic technology by the world’s leading economies’.⁴² We have already seen the ability to break classical encryption methods, which poses a significant security threat, meaning governments and organizations must prepare for the advent of quantum-resistant cryptography to protect sensitive information. Elsewhere QC can expedite drug discovery by simulating molecular interactions and predicting drug efficacy with high precision. This could lead to the development of new drugs and therapies at an accelerated pace. Quantum computers can help design new materials with exceptional properties, revolutionizing industries like electronics, aerospace, and energy storage, and enhance climate modelling, enabling more accurate and detailed simulations to understand and mitigate the effects of climate change. QC can optimize complex supply chains, leading to cost savings and reducing environmental impact, as well as revolutionize risk assessment, portfolio optimization, and fraud detection in the financial sector. Lastly, quantum machine learning can accelerate AI training and enable the development of more powerful AI models for image recognition and natural language processing. All these industries will have, and are already having, significant impacts on humanity and the planet. Public governance and oversight of them, therefore, should not be a debate. And the role of QC, both as a standalone technology, and in conjunction with these industries and sectors only heightens the need for effective public governance going forward.

4 The Rise of Quantum Computing: What’s at Stake?

As the examples described a moment ago show, QC is arguably the most exciting and sublime of all ‘disruptive’ or ‘emerging’ technologies with the potential to impact private and public governance networks and innovate public services and administrative decision-making procedures in the next decade and beyond. By harnessing the laws of quantum mechanics, QC can solve problems far in advance of classical computers and do so with much greater speed and efficiency, thus engendering degrees of quantum “advantage” and “supremacy” over classical counterparts.⁴³ This paradigm shift in the world of information processing has the potential to revolutionize numerous fields.⁴⁴ Whilst debates and discussions on QC’s power and prowess often concern impact on the material and chemical sciences and engineering, enhanced capabilities for optimisation and factorisation could equally enable the design of wholly new models of social, economic, and governmental practice.

But we must remain mindful, as with any new or emerging technology, not to overhype its potential or fail to approach it critically. That is, to ask what is at stake

⁴² WEF (2022b), p. 3.

⁴³ Atik and Jeutner (2021); Preskill (2021).

⁴⁴ WEF (2022a), p. 4.

from technologies such as QC and who stands to benefit or lose. As de Jong reminds us, ‘When looking at the case of quantum technology, we certainly see a dynamic of mystification, framing it as producing magical machines that will solve all the problems we run into with our classical methods and devices, when in fact, they promise to provide an advantage for a special class of problems.’⁴⁵ De Jong’s point is echoed by Atik who reminds us that, ‘In the end, many promising applications will yield disappointment, as quantum computers will fail to outperform other tools that are simpler yet better suited to the task.’⁴⁶

Scalability of QCs is an important factor to consider in the growth and implementation of the technology. Building large-scale QCs with thousands or millions of qubits is a formidable engineering challenge. Current quantum hardware is still in its infancy, and it will take time to develop scalable and fault-tolerant systems. Given the highly specialized and delicate (laboratory) environments presently required for QC, and the very high construction and maintenance costs, it remains to be seen what massification of the technology will look like. Cloud services that support quantum as a service (QAaaS) protocols are obvious avenues through which QC and specific simulation environments are being introduced and expanded in a wide range of contexts of (social) operation. Whilst QAaaS remains a means of accessing QCs, the experience for users is platform-like, involving well-imagined and user-friendly interfaces.⁴⁷ Perhaps unsurprisingly it is the major global technology corporations such as IBM, AWS (Amazon), and Google taking the lead in making available QC hardware for running and testing, albeit modest, quantum programs.⁴⁸

But the concentration of QC power in only a few hands is concerning, albeit a concern that is all too familiar in the information age if not before. As the WEF report highlights, ‘If the development of quantum computing moves to a few nation states and larger organizations, there is an increased risk that the most socially beneficial use cases, such as combating climate change, will be deprioritized in favour of applications that confer a competitive or geostrategic advantage to a particular group.’⁴⁹ Monopolisation (or perhaps feudalisation),⁵⁰ as a combinatorial legal, economic, and social phenomenon, is pervasive with technologies, and QCs are no different in this regard. More competition and markets in QCs and sundry applications, and the anti-trust frameworks to facilitate them are, however, not meaningful remedies to the dominance of the tech giants because, amongst other things, they do not guarantee regulation and operation that foregrounds public interest but instead market and chaotic forms of ‘rule by entrepreneur’.⁵¹

⁴⁵ De Jong (2022), p. 13.

⁴⁶ Atik (2022), p. 13.

⁴⁷ See, for example, IBM’s Qiskit site: <https://qiskit.org/>.

⁴⁸ See <https://quantumai.google/cirq/google/concepts>.

⁴⁹ WEF (2022a), p. 18.

⁵⁰ See Morozov (2022).

⁵¹ See Herian (2019).

Despite the promise of QC, several significant challenges must be addressed. QC is on the cusp of transforming industries and revolutionizing our approach to complex problems. While the technology is still in its infancy, the potential impact is undeniable. The development of QC comes with challenges, from error correction to scalability, but as these obstacles are overcome, the stakes continue to rise. The ability to harness the power of QC will determine the winners and losers commercially as well as in public sectors. Ensuring a balance is struck between those wielding the power of QC and those subject to it is crucial as the technology emerges through a wide variety of use cases and especially as it creates the types of burden described earlier. Regulation by design that, for example, operates through software written by designers and engineers,⁵² gives users and stakeholders some sense that a required balance is being struck, albeit one at high risk of non-compliance. But deferring to design methods alone will not necessarily or always address overriding human interests in opportunism and exploitation challenged through technologies. As Winner said more broadly of technologies, but which also applies to public QC implementations and a clear echo of the WEF's concerns:

The things we call “technologies” are ways of building order in our world. Many technical devices and systems important in everyday life contain possibilities for many different ways of ordering human activity. Consciously or unconsciously, deliberately or inadvertently, societies choose structures for technologies that influence how people are going to work, communicate, travel, consume, and so forth over a very long time. In the processes by which structuring decisions are made, different people are situated differently and possess unequal degrees of power as well as unequal levels of awareness.⁵³

When it comes to the responsibility placed on regulations to manage competing human interests, behaviours, and expectations, it is important to follow desire as well as need as a motivating factor. The interpersonal, replete as it is with desires and needs, ought to be, therefore, a primary concern for regulators when viewing and weighing the inputs and outputs of technologies in the interests of human stakeholders. As a body of law that places the interpersonal at the centre of legal and regulatory functions, fiduciary law, and by extension fiduciary governance, should be taken seriously as a means of addressing the interpersonal as well as managing burdens associated with emerging technologies such as QC.⁵⁴

⁵² See Almada (2023).

⁵³ Winner (1986), pp. 28–29.

⁵⁴ Evan Fox-Decent (2014) also proposes the interpersonal as characteristic of fiduciary authority. However, he draws on Kant to build this theory and situates it in the context of parental duties, which are both interesting perspectives but outside of the scope of the present discussion.

5 Fiduciary Governance

Fiduciary governance is perhaps most widely known in the world of finance and investment, particularly within the context of managing assets and resources on behalf of others. As a relative of fiduciary law, we might argue that fiduciary governance also encompasses a set of ‘subsidiary and prophylactic’⁵⁵ principles, practices, and responsibilities that guide individuals and organizations in their roles as fiduciaries.⁵⁶ Fiduciaries are entrusted with the duty to act in the best interests of those they serve, putting the welfare of the beneficiaries ahead of their own interests in a ‘service mandate’⁵⁷ that foregrounds the interpersonal and acknowledges burdens within the scope of the fiduciary’s role. As such, fiduciary duty is a cornerstone of trust, integrity, and ethical behaviour in the financial industry, but has applications further afield both actual (e.g., doctors and patients) and potential (e.g., QC engineers and end users). It is important to note that governance mandates ‘involve administration for particular purposes’, which alleviates the need found in service mandates (more traditional modes of duty and obligation found within fiduciary law) to specify beneficiaries and ‘their ascertained or ascertainable interests’.⁵⁸ This is important because it allows for a far broader application of fiduciary duties and obligations beyond the fundamental need to administer to the affairs or property of individuals. It also, I suggest, creates a way to understand how fiduciary law can address the sort of burden of knowing and know-how captured within the notion of personal administrative burden discussed above.

Fiduciary governance forms the bedrock of ethical financial management, emphasizing trust, responsibility, and loyalty in the stewardship of assets. ‘Fiduciary governance mandates reflect goals or commitments defined in relation to groups, associations, or communities’, Miller and Gold explain, and ‘are so tied to the interests of the group, community, or association that the interests of individuals cannot be disaggregated from those of the collectivity without undermining either the collectivity or the integrity of the purpose(s) posited for it.’⁵⁹ Miller and Gold continue,

purpose(s) might be framed in terms of the general public interest or in terms of general characteristics of persons that apply widely (for example, income level, educational background, etc.), such that little can be said of those intended to benefit other than that they actually or potentially fall within the scope of the mandate. A governance mandate is instrumental in bringing a well-defined collectivity into existence; the mandate, and a legal form of association with which it is connected, make organization of that kind possible, such as benefit corporations or workers’ cooperatives.⁶⁰

⁵⁵ Conaglen (2005), p. 453.

⁵⁶ Whilst the present discussion does not apply Conaglen’s fiduciary analysis in full it is important to recognize its limitations. For nuance on and criticism of Conaglen’s approach see: Lee (2007).

⁵⁷ Miller and Gold (2015), p. 517.

⁵⁸ Miller and Gold (2015), p. 517 and pp. 521–522.

⁵⁹ Miller and Gold (2015), p. 524.

⁶⁰ Miller and Gold (2015), p. 524.

The roots of fiduciary governance can be traced to ancient Rome, where the concept of “fiducia” embodied the trust between parties involved in financial transactions. However, the modern understanding of fiduciary responsibilities gained prominence in English common law during the 19th century.⁶¹ The landmark case of *Keech v Sandford*⁶² established the principle that a fiduciary must act in the best interests of the beneficiary, marking a crucial milestone in the development of fiduciary law. More recently in *Bristol and West Building Society v Mothew*, Millet LJ outlined a fiduciary as ‘someone who has undertaken to act for or on behalf of another in a particular matter in circumstances which give rise to a relationship of trust and confidence. The distinguishing obligation of a fiduciary is the obligation of loyalty.’⁶³ Hence, fiduciaries are bound by an unwavering duty of loyalty to their beneficiaries. This principle dictates that fiduciaries must prioritize the interests of those they serve over their personal interests.⁶⁴ Conflicts of interest must be disclosed and managed to ensure transparency and prevent any compromise of loyalty. And for Miller and Gold it is important to consider ‘fiduciary governance in the context of the hallmark fiduciary obligation of loyalty, contrasting the familiar ways in which fiduciaries exhibit loyalty to persons with the notion of *loyalty to purposes* that constrains execution of fiduciary governance mandates.’⁶⁵

We will return to the pressing question of accountability and enforcement for breach of obligations shortly because this must be a priority if fiduciary governance is to be taken seriously. First, however, let us consider that the standard of care that a reasonable person would employ in similar circumstances dictates the expectations on a fiduciary in any given circumstance. In the world of finance, diligent decision-making, prudent investment strategies, and ongoing monitoring of assets are essential components of fulfilling the duty of care. One of the central tenets of fiduciary governance is the obligation to follow a prudent investment standard.⁶⁶ This involves making informed and carefully considered investment decisions, diversifying portfolios to mitigate risk, and maintaining a level of diligence consistent with prevailing market conditions.⁶⁷ Defined as stewards of financial well-being, fiduciaries can be said to play a pivotal role in shaping a sustainable and ethically grounded future for financial management, a characteristic that may well prove vital in public governance settings.

Central to the present discussion is whether it is possible or desirable to extrapolate (investment) standards, encompassing monitoring, guardianship, and stewardship, to the design, engineering, and implementation of QC. This is especially

⁶¹ See, for example, *Bray v Ford* [1896] AC 44.

⁶² (1726) 25 ER 223.

⁶³ [1998] Ch 1, 18.

⁶⁴ As well as case law, the likes of the *Companies Act* 2006, ss.170-1 also enshrine fiduciary duties in statute, including as to the judgment, care, skill, and diligence of company directors.

⁶⁵ Miller and Gold (2015), p. 518.

⁶⁶ *Cowan v Scargill* [1985] Ch 270. For a further detailed analysis of this principle (which led to law reform proposals in the UK), see: Kay (2012).

⁶⁷ *Nestlé v National Westminster Bank plc* (1996) 10 TLI 113.

important, I suggest, in public service sectors where the state could (should?) intervene on behalf of citizens who may not know let alone understand QC and the ramifications of its use and, therefore, need their interests represented and protected. To relieve the burden on the citizen of having to know or understand even the basics of the technologies forming the digital infrastructure of social systems thus requires capacity and competence within the organs of state. Importantly, there must be mechanisms able to hold accountable on behalf of citizens those with capacity and competence to design, build, and implement digital infrastructure, including that driven by QC.

Given the dominance of private commercial actors in present digital infrastructure, there is little reason to believe QC will be any different. Indeed, all the evidence points towards the dominance of legacy tech corporations such as IBM, Google, and Amazon taking the lead in scaling-up QC. Therefore, it is necessary for the successful implementation of QC to negotiate a private/public relationship going forward. And in fiduciary governance we find precedents for managing such relationships. As Miller and Gold explain:

A governance mandate can be established for any of a number of private or public purposes, and the purposes in question can be posited by an individual acting in a private or official capacity, or by a group of individuals acting likewise as private persons or members of a public body. Purposes specified for a governance mandate may reflect interests in forms of common property (for example, homeowners' associations in cooperatives), or personal or community interests (for example, the incorporation of a leisure club). Individuals may also establish a governance mandate for purely public purposes relating to some aspect of general public welfare.⁶⁸

The advent of AI has introduced new challenges for governments, and the same applies to QC. Given that QC is in a more nascent stage of its development and evolution than AI, intervening sooner rather than later to ensure, for example, the public welfare Miller and Gold describe, is both possible and critical. The need to navigate complex algorithms, assess the risks associated with digital assets, and ensure cybersecurity poses novel hurdles in maintaining the duty of care, and, as we have seen, because QC operates in ways unfamiliar and even alien to human experience, the necessity to maintain a resolute and meaningful duty of care standard is that much more pressing.

Contemporary fiduciary duties acknowledge the need for fiduciaries to handle vast amounts of sensitive financial and personal data. Safeguarding this information from cyber threats and ensuring compliance with data privacy regulations are already critical components of fiduciary governance in the digital age. But existing safeguarding strategies are, again, cast into doubt by the potential for a profound difference between classical and quantum computing methods and outputs, which means a continuous level of monitoring. As investment managers must continuously monitor the performance of the investments in their clients' portfolios, regularly reviewing and adjusting the portfolio to ensure it remains in line with the client's

⁶⁸ Miller and Gold (2015), p. 525.

objectives, the same principle, it is argued, can and should be applied to the QC ecosystem.

Fiduciary governance is not without its challenges, and navigating these challenges is essential for fiduciaries to fulfil their duties effectively. Conflicts of interest can arise in various fiduciary roles and there is no reason to believe the same would not apply in terms of QC if stakeholders are not held to account. For example, investment advisers may be tempted to recommend products that offer them higher commissions but are not in the client's best interest. Likewise, implementation of QC in a public service sector may be directed at private commercial gain rather than necessary or clear public benefit. As Miller and Gold maintain:

In fiduciary law, as in other areas of private law, interpersonal principles of liability make one person accountable to another in respect of interpersonal conduct. Generally speaking, interpersonal accountability is effectuated through correlative rights and duties. Just as fiduciary service mandates and relationships serve to facilitate extension of the personality of persons through delegation of powers to promote the ends of particular *persons*, governance mandates enable extension of the personality of persons through delegation of powers to promote particular *abstract purposes*.⁶⁹

Addressing and mitigating conflicts is a significant challenge, and fiduciaries must navigate a complex web of regulations and laws that hold them to account. Fiduciary governance helps mitigate risks by requiring diversification, prudent decision-making, and thorough due diligence. Fiduciaries are traditionally held accountable for their actions by virtue of the rules and principle outlined earlier, and beneficiaries have recourse to the courts if they believe fiduciary duties have been breached. Fiduciary accountability under governance-type mandates is, therefore, '*structurally* institutional in that it is served by *standalone duties* that may be enforced by any of a number of persons or entities occupying a monitoring and enforcement role in respect of the mandate.'⁷⁰ Further, 'enforcement powers may be wielded through a private right of action or a derivative action for a personified entity. Alternatively, or additionally, when enforcement powers are held by public officials or government agencies, accountability may be achieved through direct administrative or regulatory action.'⁷¹

To ensure enforcement is credible, therefore, robust processes and forums for adjudication are required. More traditional adjudication of breaches of fiduciary duty, for instance, first require the fiduciary to give an account of their actions, including the nature, type, and risk profile of investments, to court or in some circumstances a panel or mediator conducting alternative dispute resolution (ADR). Remedies flowing from a breach may include personal liability for the fiduciary in the form of restitution or compensation. Whilst borrowing from what are in large part legal rules drafted in response to private finance, there is no reason an enforcement and remedial framework such as this cannot work for fiduciary accountability

⁶⁹ Miller and Gold (2015), p. 544 and p. 551.

⁷⁰ Miller and Gold (2015), p. 554.

⁷¹ Miller and Gold (2015), p. 556.

under governance-type mandates, as, indeed, Miller and Gold argue. In the case of public benefit requirements, the position taken must be one of testing resultant harms or consequences to the community to ensure justice in the public interest. Consequently, a robust fiduciary governance framework is one that disincentivises self-interest (e.g., at a personal, corporate, or institutional level) and opportunism that lead inexorably to a breach in obligations, duties, and responsibilities. The aim from this interplay between *ex ante* and *ex post* governance standards is to ensure stakeholders in QC, and those such as government agencies that seek to implement the technology in public domains, tailor their ambitions to the needs and priorities of the community.

6 Quantum Computing with Public Governance

As we have seen, QC offers the potential to solve problems currently unfeasible for classical computers, including optimizing resource allocation, simulating complex systems, enhancing encryption, and advancing machine learning algorithms.⁷² While the potential benefits of QC in public governance are substantial, there are several challenges to overcome. First and foremost is the issue of hardware scalability. Quantum computers are still in the experimental phase, and building practical, large-scale quantum machines remains a significant technical challenge. Building quantum expertise within government agencies is also crucial for maximizing the benefits of QC in public governance. Additionally, governments need to address ethical and privacy concerns associated with the collection and analysis of vast amounts of data. Not least because transparent and responsible data governance policies are essential to maintain public trust and uphold fiduciary duties.

In public governance contexts, QC's computational prowess can be harnessed in various ways and interweaved throughout a meshwork of societal and civic service mandates, examples of which we will look at below. Against this backdrop of emerging contexts and use cases, it is vital to remember that QC requires significant oversight to protect the interests of citizens and hold accountable those that breach duties and obligations within the scope and competence of public governance, both private and public actors alike. Fiduciary governance as described above ensures the monitoring, stewardship, and guardianship of the technology and its use within civic domains for the public benefit. Further, fiduciary governance can operate as a supplementary means alongside existing regulations, including data and consumer protections, to manage human behaviours, interests, and expectations relating to the implementation and use of QC in the service of civic aims.

A primary responsibility of government is to allocate resources efficiently, ensuring that taxpayer funds are collected and utilized for the benefit of all citizens. QC can play a pivotal role in optimizing resource allocation along these lines. For

⁷² See, for example, D-Wave (2021).

example, government agencies responsible for budgeting, procurement, and infrastructure development in healthcare settings (e.g., the NHS in the UK) can leverage targeted quantum algorithms to streamline transportation routes and optimize supply chain logistics, ensuring that essential goods reach their destinations more efficiently. These improvements not only save taxpayer money but should enhance the overall quality of public services by optimizing budget allocation to maximize resource utilization while minimizing waste. QC also has the potential to revolutionize public policy formulation and analysis based on the same advanced capabilities discussed above. Therefore, with the ability to process vast amounts of data and run complex simulations, quantum computers could help governments model the long-term impacts of policies and regulations more accurately. This enables data-driven decision-making that is grounded in empirical evidence. Hence, QC can be used to model the economic impact of different taxation policies suggested a moment ago, enabling governments to make informed choices that optimize revenue while minimizing the burden on citizens. This not only aligns with fiduciary responsibility by meeting the needs and addressing the burden on citizens, but it also promotes transparency and accountability in public governance.

Data security is a concern of utmost importance for governments. As discussed above, the quantum threat to classical encryption algorithms has long been recognized, and with the advent of QC, traditional or classical cryptographic methods are at risk of being cracked, potentially exposing sensitive government data.⁷³ However, QC also offers solutions in the form of quantum-resistant encryption and algorithms that governments can use to secure data against quantum attacks. This is crucial for protecting, amongst other things, national security information, personal citizen data, and sensitive financial records. Quantum-safe encryption methods are an essential component of quantum-ready public governance as government agencies and departments with vast amounts of sensitive data are prime targets for cyberattacks. QC can be employed to enhance cybersecurity through the creation of quantum-resistant cryptographic methods, including a variety of digital signatures.⁷⁴ These types of post-quantum encryption technique are designed to withstand quantum attacks, safeguarding government data against unauthorized access and breaches. Moreover, QC can be used to detect and respond to cyber threats in real-time. Machine learning algorithms running on quantum hardware can analyse massive datasets for signs of intrusions or vulnerabilities, allowing for proactive cybersecurity measures that support and maintain the obligations and duties of government and the interests of citizens.

⁷³ See also: NCSC (2020).

⁷⁴ See, for example, NIST (2024).

7 Conclusion

This contribution has argued that there is an increasingly heavy techno-social burden on citizens. Accordingly, there is a need for individuals and communities to have belief or faith—what I refer to here as *quantum fide* (as much as faith)—in the stakeholders of technologies such as quantum hardware and systems because knowledge gaps not only persist but widen. In short, it is increasingly hard to know or to understand how the technologies that shape our lives work and what is at stake when using or being subjected to them. The burden is not one that can be easily solved by education or digital literacy programmes, nor in fact should it be. This is not a position that necessarily or inevitably leads to technocracy; experts certainly have an important role to play and should not be dismissed, but they need not be the ones that determine policy or define rules. Social democracy and the governments it can foster must ensure oversight and accountability of stakeholders. This can both relieve the techno-social burden and show that technologies, insofar as they are aimed at, have consequences for, and impact citizens, do indeed hold real promise and advantages for all, equitably. Hence, building on the WEF's Quantum Computing Governance Principles, it is imperative for the likes of the UK government's National Quantum Strategy to take seriously the stated need to find a balance between innovation and ethical considerations.⁷⁵

Enabling or ensuring public faith in governments to implement QC in the best interests of citizens is a vexed issue, one that requires a robust yet flexible means of accountability able to test boundaries and enforce breaches. I have speculated on whether fiduciary obligations, duties, and responsibilities that coalesce to form meaningful and enforceable techno-social governance frameworks for the management of a variety of QC implementations is an option to move forward with QC orientated public innovation programmes. From optimizing resource allocation to enhancing data security, accelerating scientific research, and revolutionizing public policy, QC can revolutionize the way governments serve their citizens if the technology is managed for the benefit of the many. Successful integration of QC into public governance requires careful stewardship as well as planning and investment. It is imperative that governments understand the potential of QC and work towards harnessing its power to reap the many benefits on the horizon, to better address the complex challenges of the modern world and ensure the well-being of their citizens.

Acknowledgments I would like to thank the organizers and participants of the workshop in Utrecht on 11 January 2024 for their useful comments and feedback on this chapter.

⁷⁵ Department for Science, Innovation & Technology (2023), pp. 47–48.

References

- Almada M (2023) Regulation by design and the governance of technological futures. *Eur J Risk Regul* 14:697–709
- Atik J (2022) Quantum computing and the legal imagination. *SciTech Lawyer* 18(3):12–16
- Atik J, Jeutner V (2021) Quantum computing and computational law. *Law, Innov Technol* 13(2):302–324
- Bennett CH, Brassard G, Mermin ND (1992) Quantum cryptography without Bell's theorem. *Phys Rev Lett* 68:557–560
- Carroll S (2019) Something deeply hidden: quantum worlds and the emergence of spacetime. *One World*
- Cole M (22 August 2022) What's wrong with technocracy?, *Boston Review*. <https://www.boston-review.net/articles/whats-wrong-with-technocracy/>
- Conaglen M (2005) The nature and function of fiduciary loyalty. *LQR* 121(July):452–480
- Davis D, Kim A (17 August 2023) Quantum computing could threaten Blockchain, Crypto. *Bloomberg Law*. <https://news.bloomberglaw.com/us-law-week/quantum-computing-could-threaten-blockchain-crypto>
- de Jong E (2022) Own the unknown: an anticipatory approach to prepare society for the quantum age. *Digit Soc* 1(15):1–23
- Department for Digital, Culture, Media & Sport (2022) UK digital strategy. Department for Digital, Culture, Media & Sport. https://assets.publishing.service.gov.uk/media/62c7050fd3bf7f3008697a7e/UK_Digital_Strategy_web_accessible.pdf
- Department for Science, Innovation & Technology (2023) National Quantum Strategy. Department for Science, Innovation & Technology. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1142942/national_quantum_strategy.pdf
- D-Wave (5 October 2021) D-Wave customer applications. D-Wave. <https://www.youtube.com/watch?v=5n CvPOVODoQ>
- Edelman J (2014) The role of status in the law of obligations. In: Gold A, Miller PB (eds) *Philosophical foundation of fiduciary law*. Oxford University Press, pp 21–38
- Feynman R (1998) *Six easy pieces: the fundamentals of physics explained*. Penguin
- Fineman MA (2008) The vulnerable subject: anchoring equality in the human condition. *Yale J Law Femin* 20(1):1–23
- Foucault M (2002) *Power: essential works of Foucault 1954–1984*. Penguin
- Fox-Decent E (2014) Fiduciary authority and the service conception. In: Gold A, Miller PB (eds) *Philosophical foundation of fiduciary law*. Oxford University Press, pp 363–387
- Galloway A (2004) *Protocol: how control exists after decentralization*. MIT Press
- Giles M (29 January 2019). Explainer; what is a quantum computer? *MIT Technology Review*. <https://www.technologyreview.com/2019/01/29/66141/what-is-quantum-computing/>
- Giurge LM, Whillans AV, West C (2020) Why time poverty matters for individuals, organisations and nations. *Nat Human Behav* 4:993–1003
- Greene B (2000) *The elegant universe: superstrings, hidden dimensions, and the quest for the ultimate theory*. Vintage
- Herian R (2019) *Regulating blockchain: critical perspectives in law and technology*. Routledge
- Herian R (2021) *Capitalism and the equity fetish: desire, property, justice*. Palgrave Macmillan
- Herian R (2022) *The conscience of thomas more: an introduction to equity in modernity*. Heythrop J 63:64–75
- Jeutner V (2021) The quantum imperative: addressing the legal dimension of quantum computers. *Morals Mach* 1(1):52–59
- Kay J (2012) *The Kay Review of UK Equity Markets and Long-Term Decision Making: Final Report*. <https://assets.publishing.service.gov.uk/media/5a7ccdbed915d6b29fa8bea/bis-12-917-kay-review-of-equity-markets-final-report.pdf>
- Knapp S, Michaels WB (1982) Against theory. *Crit Inq* 8(4):723–742

- Kop M, Aboy M, de Jong E, Gasser U, Minssen T, Cohen IG, Brongersma M, Quintel T, Floridi L, Laflamme R (2023) Towards responsible quantum technology: safeguarding, engaging and advancing quantum R&D. *UC Law Sci Technol J* 15(1):63–94
- Lee R (2007) In search of the nature and function of fiduciary loyalty: some observations on Conaglen's analysis. *Oxford J Leg Stud* 27(2):327–338
- Miller PB, Gold AS (2015) Fiduciary governance. *William & Mary Law Rev* 57(2):513–586
- Morozov E (2022) Critique of techno-feudal reason. *New Left Rev* 133/134:145–160
- NCSC (2020) Quantum security technologies. NCSC. <https://www.ncsc.gov.uk/whitepaper/quantum-security-technologies>
- NIST (2024) Post-quantum cryptography: digital signature schemes. NIST. <https://csrc.nist.gov/projects/pqc-dig-sig/round-1-additional-signatures>
- Perrier E (2022) The quantum governance stack: models of governance for quantum information technologies. *Digit Soc* 1(22):1–48
- Preskill J (2021) Quantum computing 40 years later. Institute for Quantum Information and Matter. California Institute of Technology
- Tinmaz H, Lee YT, Fanea-Ivanovici M, Baber H (2022) A systematic review on digital literacy. *Smart Learn Environ* 9(21):1–18
- Winner L (1986) *The whale and the reactor*. University of Chicago Press
- World Economic Forum (2022a) Quantum computing governance principles: insight report. World Economic Forum. https://www3.weforum.org/docs/WEF_Quantum_Computing_2022.pdf
- World Economic Forum. (2022b). State of quantum computing: building a quantum economy: insight report. <https://www.weforum.org/publications/state-of-quantum-computing-building-a-quantum-economy/>
- Yao Q, Zhang B, Luo Y, Huang H (2014) No-cloning theorem, Leibniz's principle, and the notion of uniqueness. *Open Access Libr J* 1:1–5

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.



The Rule of Tech? Submitting Legislation to Digital Constraints



Sabrina Kutscher and Johan Wolswinkel

Abstract While ‘by design’ approaches are increasingly advocated to ensure technology being compliant with legislation, the opposite challenge of adopting legislation that is compliant with digital demands, is also gaining ground. Several jurisdictions are experimenting with some form of ‘digital-ready’ legislation, which aims at ensuring a smooth implementation of legislation through digital systems, but which also transforms the legislative process itself. This chapter explores and compares several instances of digital-ready legislation and assesses the consequences thereof for the technocracy-democracy divide and the virtues commonly associated with the legislative process. It concludes that legislation is increasingly subject to the ‘rule of tech’ next to the ‘rule of law’, while the virtues traditionally associated with the legislative process have not become irrelevant.

1 Introduction

Emerging technologies are assuming an increasingly prominent role in the realm of public governance, offering the potential to enhance the accessibility, efficiency, and expeditious delivery of public services.¹ Although the implementation of automated public services is promising, it also comes with challenges. Some challenges have been met with so-called ‘by-design’ approaches, which aim to ensure *legal*

¹Manoharan et al. (2023), pp. 265–266.

S. Kutscher

Tilburg Institute for Law, Technology and Society, Tilburg Law School, Tilburg University,
Tilburg, The Netherlands

e-mail: s.k.kutscher@tilburguniversity.edu

J. Wolswinkel (✉)

Public Law and Governance, Tilburg Law School, Tilburg University,
Tilburg, The Netherlands

e-mail: c.j.wolswinkel@tilburguniversity.edu

compliance already at the stage of designing digital tools.² This, however, constitutes only one side of the coin, as the interactions between the legal and the digital sphere are bi-directional. The other side, which is often overlooked, is the issue of *digital* compliance: (how) is legislation drafted (by design) to meet digital constraints and needs, for instance as regards the automated implementation of legislation? Whereas legislation is said to be bound only by the rule of law, it could be equally bound by the ‘rule of tech’ in the era of digital public governance.

Thus, instead of just asking how the digital realm complies with legal principles and values, it is equally important to examine how the legal dimension must submit to the digital architecture. This question is especially pertinent, as, at present, several countries are already experimenting with this other side of the coin. Denmark, for example, has adopted the idea of “digital-ready legislation”, whereas New Zealand has facilitated pilot programs with so-called “machine-consumable legislation”, and the Netherlands has been working on “digital-workable legislation”. These initiatives could be coined as part of an ongoing ‘rules as code’ movement, which the OECD characterizes as “a fundamental change in the status quo of rulemaking”.³

This chapter will examine the interaction between these two realms in the design of legislative processes by pursuing a dual objective. First, we aim to further develop the concept of ‘digital-ready legislation’⁴ to expand knowledge on this subject and increase its visibility by exploring the landscape and concrete instances of digital-ready legislation. Second, we offer some reflections on what this trend means for the development and understanding of public values central to legislative processes, such as democracy, transparency, and accountability. This will be done by conducting a policy document analysis on digital-ready legislation in three jurisdictions (Denmark, Aotearoa (New Zealand) and the Netherlands,) and by comparing and synthesizing the results thereof.

This article is structured as follows. First, we discuss previous work on the relative virtues of legislation and relate the concept of legislative supremacy to the technocracy/democracy debate (Sect. 2). Next, we further explain our comparative method (Sect. 3) and describe three regulatory initiatives on making legislation more “digital-ready” or “machine-consumable” (Sect. 4). Based on these descriptions, we compare the different approaches to assess which different constellations of reconciling technocracy with democracy can be distinguished (Sect. 5). In our concluding section, we evaluate the development towards digital-ready legislation and propose avenues for further research.

² Diver (2024), pp. 2–3.

³ Mohun and Roberts (2020), p. 16.

⁴ Unless otherwise provided, we will refer to ‘digital-ready legislation’ and ‘machine-consumable legislation’ interchangeably.

2 Theoretical Background

2.1 Legislative Supremacy

To evaluate the relative merits of legislation over technology (or vice versa), it is necessary to first assess the virtues ascribed to legislation as such. Usually, these virtues of legislation are identified by comparing legislation with other sources of law. In the context of the separation of powers, much attention is paid to the hierarchy either between the legislative and the judicial branch of government or between the legislative and the executive branch of government. In that regard, the concept of *legislative supremacy* (sometimes also referred to as parliamentary sovereignty) stresses the domination of legislation over other sources of law. Several (sets of) virtues of the legislative process have been put forward favoring the legislature as lawmaker, which we will discuss briefly below, following Daly.⁵

First, the legislative process has traditionally been linked with *democracy*. Democracy entails (i) that political power is exercised by (and only by) the people, and (ii) that citizens are equal in the exercise of this political power.⁶ Thus, democracy is closely related to the principle of procedural equality.⁷ Since no citizen's view is privileged or prejudiced 'upfront' in the decision-making process, every citizen has an equal vote in this legislative process.⁸

A *second* set of virtues relates to the *discursive* qualities of the legislative process, which is associated with practices such as compromise, deliberation, and debate. While it has been argued that these discursive or reason-giving qualities of the legislative process might be too contextually contingent, or insufficiently generalizable to offer any resilient justification for legislative supremacy, at least they constitute an additional argument next to procedural equality in support of legislative supremacy.⁹

A *third* set of virtues, which applies to the hierarchy between legislation and case-law, is the *purposive generality* of legislation. Unlike the two sets of virtues discussed above, purposive generality refers to the outcome (or 'form') of legislation rather than to the process of enacting legislation. Legislation, by its very nature, is drafted to cover general categories and principles rather than specific persons and situations. As such, legislatures therefore do not rule directly over individuals, but only with reference to abstract categories.¹⁰ Case law, by contrast, even though its effects sometimes exceed the individual case, is in itself not aimed at creating general rules.

⁵Daly (2020), pp. 807–830.

⁶Beetham (1999), pp. 1–30.

⁷Daly (2017), pp. 501–531.

⁸Daly (2017), p. 503.

⁹Daly (2017), p. 504.

¹⁰Daly (2017), pp. 510–512.

A *fourth* set of virtues relates to *transparency*, *accountability*, and *publicity*. Transparency is often considered a necessary condition for accountability, with publicity as a species of transparency, referring to the idea that public norms should be known to and accessible by the citizens who are subjected to them.¹¹ Transparency has a dual relationship with the virtues assigned to legislation. On the one hand, it is associated with the legislative process. In that regard, transparency of the legislative process can be considered a corollary of the democratic process involved, thereby also strengthening accountability of that process. On the other hand, transparency can also be linked with legislative outcome (or form) rather than process). Since legislation is meant to be generally applicable, its general wordings contribute to clarity and comprehensibility of legislation. Thus, transparency supports the purposive generality of legislation.¹²

While these four sets of virtues have been put forward in favoring legislation over case-law, the debate about legislative supremacy from the perspective of the executive branch usually centers around the first set of virtues related to democracy.¹³ The central point of critique towards delegation of legislative power to government is the lack of sufficient democratic legitimation.¹⁴ However, evaluating delegated law-making from the perspective of all four sets of virtues gives a more nuanced picture. Admittedly, delegated law-making might also suffer from a lack of deliberation and a lack of transparency due to the absence of parliamentary debates. At the same time, delegated law-making shares the purposive generality which is inherent to parliamentary legislation, as this follows from the outcome (form) rather than the process of law-making.

It is important to note that the virtues of legislation discussed above are characteristic for Western liberal democracies, but do not necessarily apply universally. What is more, the relative weight of each of these values can also differ across these democracies. This is exemplified, for instance, by whether a court has competence to evaluate the constitutionality of legislation. Nonetheless, these virtues of legislation, irrespective of their exact weight, can be useful parameters to contrast legislation with other sources of law.

2.2 *Democracy Versus Technology*

Even if legislation itself is supreme over delegated law-making or case-law because of its relative virtues, the question arises whether the legislature itself is bound by other constraints. It has been argued that legislative supremacy itself is subject to the

¹¹ Daly (2017), p. 525.

¹² Daly (2020), p. 811.

¹³ As far as single-case decision making is concerned, the arguments favoring legislative supremacy over case law might apply *mutatis mutandis*.

¹⁴ Pünder (2009), pp. 353–378.

rule of law, thereby limiting the sovereignty of parliament in enacting legislation.¹⁵ With the rise of new technologies in the implementation of legislation, the question arises whether legislative supremacy might also be subject to another rule, the “rule of tech”, and therefore has to abide by the rules of the digital architecture.¹⁶ De Gregorio, for example, suggests in this regard that the introduction of AI technologies results in an additional normative layer where norms and rules originate from the technology itself and may compete with the rule of law.¹⁷ Others point to the role of private actors and the consolidation of power within their technical infrastructures that enable these actors to shape law.¹⁸ However, the focus of this chapter is on the notable rise of new technologies and the challenges it brings relating to the idea of legislative supremacy. Legislative supremacy is challenged for two reasons: (1) there is often the argument that legislation does not sufficiently accommodate technical developments and innovations due to its slow pace,¹⁹ and (2) the complexity of technological innovation requires expertise to effectively regulate it and therefore reliance of the legislature on technocrats.²⁰ Hereafter, this section will discuss both of these considerations, as they are important in assessing the ways in which legal and digital constraints can be reconciled as well as the effects this could have on the virtues that are fostered through the concept of legislative supremacy.

First, in the discipline of systems safety, the misalignment between developments and mechanisms of, on the one hand, technology and, on the other hand, control thereof via, for example, governance instruments, is referred to as “asynchronous evolution”. Although this asynchronicity can move in both directions, this phenomenon commonly describes situations in which systems, often in the context of technology, continuously evolve by adapting and reacting to environmental changes.²¹ While these systems evolve and introduce new challenges, issues, and opportunities, it can often be observed that the regulatory or legal realm remains static or will change at a delayed pace. This is often seen as a hindrance to innovation, as the inability of law to keep up with regulation of technology creates legal uncertainty.²² One way to counteract this phenomenon of asynchronous evolution, which is becoming more prevalent due to the increasing pace at which technological innovation is moving forward, is the notion of “agile legislation”. Agile legislation describes a new approach to law whereby it is recognized that environments can change quickly and require fast legal responses. This means that emphasis is put on a maintenance-centered approach in the adaptability of law to changing technological

¹⁵ See, in particular, Chapter 7 ‘The Rule of Law and Parliamentary Sovereignty’ in Allan and Allan (2001).

¹⁶ See Suzor (2019).

¹⁷ De Gregorio (2023), p. 55.

¹⁸ Terzis (2024), pp. 1, 15.

¹⁹ See Kaal (2016).

²⁰ Gilley (2017), p. 11.

²¹ Leveson (2012), pp. 85, 95, 514–516.

²² Fenwick et al. (2017), p. 573; Gaudet and Marchant (2011), p. 167.

environments by focusing on generating flexible and coordinated legal instruments and legal environments.²³ In other words, “it refers to a continuing state of readiness for action, but without an identified purpose.”²⁴

Second, with technology introducing more complexities in terms of regulation, and with countries moving towards legislation that is ‘compliant by design’ or agile,²⁵ concerns are also voiced whether these new developments may overshadow the concept of legislative supremacy.²⁶ Although attempts are made to adapt law-making processes and legislative instruments to the digital environment, the legislative branch may be limited in its procedures and work due to lack of technical expertise. In other words, adapting law-making to the digital environment may create dependencies on technical experts.²⁷ This tension between, on the one hand, legislative supremacy, which represents an important safeguard within democratic societies, and, on the other hand, technocracy, referring to the involvement of unelected experts into legislative processes, is not a new phenomenon as Gilley shows.²⁸ It has always been the case that new innovations that are introduced to society require the input of experts in effectively regulating the uses and effects. However, the question remains as to whether technocratic influence circumvents important virtues of the legislative process, including democratic accountability and transparency, and whether this could affect democratic virtues negatively.²⁹ Therefore, this tension between technocracy and democracy (in particular legislative supremacy) will be revisited to assess to what extent this creates issues with regard to democratic virtues or whether these two notions can somehow be reconciled.

3 Method

In the following, we will report on three jurisdictions pushing towards some form of ‘digital-ready-legislation’. These three jurisdictions are Denmark, Aotearoa (New Zealand) and the Netherlands, as they are all, to a greater or lesser extent, frontrunners when it comes to ‘rules as code’. Although the OECD report on rules as code also mentions initiatives in other jurisdictions,³⁰ we have chosen those jurisdictions

²³ Boer and Sileno (2013), pp. 2, 5.

²⁴ Boer and Sileno (2013), pp. 5–6.

²⁵ The wording ‘by design’ is not always used in the same meaning. Margaret Hogan, for example, contrasts a design-driven approach of law (legislation) with a tech-driven approach thereof, by defining ‘legal design’ as a human-centered approach of law, see Hogan (2015), Chapter 1.

²⁶ See Rangone (2023).

²⁷ Fenwick et al. (2017), p. 573; Sidorenko and von Arx (2020), p. 30.

²⁸ See Gilley (2017).

²⁹ Gilley (2017), pp. 10–11.

³⁰ Mohun and Roberts (2020), p. 61, referring to initiatives in Germany and Australia. For more information on rules as code in Australia, refer to <https://www.govcms.gov.au/rac>.

(Denmark and Aotearoa) that are also part of the Digital Nations initiative, a partnership of ten countries considering themselves as leading in the transformation of digital government for the benefit of citizens.³¹ In order to enrich the existing landscape of ‘rules as code’ initiatives, we have added a third jurisdiction to our comparative analysis that is neither referred to in the OECD report nor a member of the Digital Nations, but has also undertaken a ‘rules as code’ initiative. This third jurisdiction is the Netherlands, which has expressed its ambition to become a frontrunner in the development and use of responsible digital technology.³²

For each of these jurisdictions, we have collected and analyzed publicly available policy documents which are specifically targeted at digital-ready legislation. After presenting this material for each jurisdiction separately (Sect. 4), we will compare our findings for the different jurisdictions on the basis of the democracy/technocracy debate and the virtues of the legislative process (Sect. 5). Since the primary aim of this comparison is to explore the landscape and concrete instances of digital-ready legislation, the analysis is not targeted at a very close assessment of the (legal) similarities and differences between the jurisdictions under review, as is common in comparative legal research.³³ Instead, while paying attention to similarities and differences in terms of wording, principles and approaches (Sects. 5.1 and 5.2), this comparison ultimately aims to facilitate an evaluation of digital-ready legislation as an overarching concept from the perspective of the different virtues associated with legislation (Sect. 5.3).

4 Results: Variety in Digital-Ready Legislation

Considering the ever-increasing penetration of emergent technologies into various aspects of social life, it is logical that slowly attempts are made for digital technologies to also enter the legislative realm. Indeed, jurisdictions such as Denmark, New Zealand³⁴ or the Netherlands, are experimenting with merging legislative instruments and digital technologies. Although using digital technologies in the realm of law is not completely new, as some governments have been publishing their legislation in machine-readable versions, the new propositions regarding digital transformation made by these frontrunners far exceed the mere publishing of legislation that can be digested by computers³⁵ In general, whereas machine-readable legislation

³¹ See Digital Nations (2020).

³² Interestingly, especially for this concrete goal, target indicators still have to be developed see Government of the Netherlands (2022), p. 26.

³³ Refer to the 2011 article on comparative legal research, pp. 61–62.

³⁴ New Zealand now also refers to itself as “Aotearoa” which is the Māori name for their country. Therefore, this contribution will alternate between the English and Māori name, see Department of Internal Affairs (2020).

³⁵ Epifanova et al. (2020), pp. 972–973; Francart et al. (2019), p. 137; Ma and Wilson (2021), pp. 115–116; Office of the Press Secretary (2013).

allows individuals to search legislative texts for keywords or to identify the document type according to how it was marked up, digital-ready or machine-consumable legislation implies that the *meaning* of the legislative text is encoded. As such, it would be possible for a computer program to apply the law pursuant to its legislative intent.³⁶ In this section, we will elaborate more on what exactly these three countries have experimented with so far and the different approaches that were taken.

4.1 Denmark

Denmark has always been considered a frontrunner in digitalization of government. Already in 2011, the Agency for Digital Government was launched with the task of implementing policies aimed at expanding the digital public sector as well as the Danish digital service infrastructure.³⁷ This implementation has taken place over the last decade in several ways, ranging from digital infrastructure and information security to user-centered digital public services.³⁸ Currently, the Agency for Digital Government is part of the Ministry of Digital Government and Gender Equality.

As part of this growing scheme of shifting public services and decision-making into the digital sphere, one project is the move towards so-called ‘digital-ready legislation’. In January 2018, the political parties in the Danish parliament concluded an agreement pushing for the commitment of making all Danish legislation “digital-ready”. The agreement states that such legislation “will weed out unnecessary and complex rules and ensure that new rules are easy to understand and translate into secure and user-friendly digital solutions”.³⁹ Inherent to this legislative move is the aim of enhancing the digitalization of public administration by recalibrating the focus of law-making. Therefore, the political agreement concludes that from July 2018 all legislation must be digital by default to ensure a smooth digital transition within the public law sector. This means that opportunities to simplify rules and prevent ambiguity within the drafting process should be realized to expand the use of automated decision-making as well as to increase accessibility of legislation among citizens.⁴⁰ Whereas the political agreement itself stated that in 2020 the parties would assess whether the legislation is sufficiently digital by default and, if so, discuss further initiatives for supporting digital-ready legislation, the Danish government announced in its National Strategy for Digitalisation, adopted in 2022, a continuing focus on digital-ready legislation: “Legislation must be easy to understand and simple to manage digitally. The work to prepare legislation for digitization must therefore be continued and strengthened. This will ensure that legislation

³⁶ Waddington (2019), pp. 25–27.

³⁷ See Agency for Digital Government (n.d.).

³⁸ Plesner and Justesen (2022), p. 155.

³⁹ See Agency for Digital Government (2018a, b).

⁴⁰ See Agency for Digital Government (2018b); Plesner and Justesen (2022), pp. 147–148.

Table 1 Seven principles for digital-ready legislation

Principle 1	Simple, clear rules
Principle 2	Digital communication
Principle 3	Possibility of automated case processing
Principle 4	Consistency across authorities—uniform concepts and reuse of data
Principle 5	Safe, secure data handling
Principle 6	Use of public infrastructure
Principle 7	Prevention of fraud and errors

can be translated into digital solutions in an efficient, secure and ethically responsible way.”

The push towards digital-ready legislation has resulted into seven principles of digital-ready legislation, which have been formulated in the political agreement and which have been further developed by the Danish Agency for Digital Government.⁴¹

In the context of legislative drafting, there are three principles of digital-ready legislation that seem particularly relevant (see bold in Table 1). Principle 1 on simple, clear rules is one indicator for the intent of ensuring that legislation becomes more adjustable to digital constraints and requirements. Although clear and simple rules are also meant to make legislation more comprehensible for citizens, there is an emphasis on reducing ambiguity and reinforcing both consistency and logical reasoning. Principle 1 also includes the overall avoidance of exceptions and special schemes where possible in favor of simplicity and clarity.⁴² This principle therefore signifies the requirement of drafting legislative texts as clear and concise as possible to facilitate easy translation of the text into code.

Principle 3 on the possibility of automated case processing can be considered as another signifier by which legislation is made more compliant with the digital infrastructure. This becomes apparent in the repeated insistence on opting for objective criteria when possible as these are considered “a prerequisite for automation of case processing.”⁴³ Moreover, Principle 3 further suggests the importance of scrutiny when it concerns the integration of discretionary powers. Although it is acknowledged that exercise of discretion can constitute an important safeguard to consider all relevant conditions in situations that demand a more sensitive assessment, Principle 3 simultaneously calls for evaluating whether discretionary powers include elements that can be changed into objective criteria.⁴⁴ Again, this signals a shift of

⁴¹ See Agency for Digital Government (2018a).

⁴² See Agency for Digital Government (2018a).

⁴³ See Agency for Digital Government (2018a).

⁴⁴ See Agency for Digital Government (2018a).

legislative process towards drafting more standardized, objective procedures that take into account the digital demands of technology.

Lastly, Principle 4 displays an effort towards digital compliance by highlighting the need for consistency and uniformity regarding concepts and reuse of data. Key words such as “cohesion”, “uniformity”, or “consistency” hint towards standardization. Where possible, concepts should be used in a uniform way throughout different pieces of legislation, thereby subjecting new legislation also to legislative choices made in other legislation.

4.2 Aotearoa (New Zealand)

Aotearoa is another country which has looked at the merge of law and code early on. This approach was a project supported by the Service Innovation Lab, which was set up in 2017 as an “all-of-government lab”, hosted by the Department of Internal Affairs as the government’s functional lead agency for digital development.⁴⁵ Thus, also in the Aotearoa context, the initiative on machine-consumable legislation can be positioned within a broader effort of the Aotearoa government in digitalizing public services. Mid-2020, the Services Innovation Lab was closed due to the formation of a new Digital Public Service branch within the Department of Internal Affairs. This new branch should deliver on the government’s new Strategy for a Digital Public Service, which highlights the goal of achieving better outcomes by creating and relying on digital public services to be more responsive, agile, and adaptive. In particular, the idea is that other (existing) governmental agencies will adopt and continue with the prototype work developed within the Services Innovation Lab, where necessary supported by the Digital Public Service branch.⁴⁶

One of the projects the Service Innovation Lab started was the “Better rules, better outcomes” initiative in January 2018.⁴⁷ By way of a three-week discovery sprint involving cross-agency and multidisciplinary stakeholders on two existing pieces of legislation,⁴⁸ the Service Innovation Lab explored the opportunities and challenges of so-called “machine-consumable legislation”. According to the resulting report, “machine consumable” legislation is understood as “having particular types of rules available in a code or code-like form that software can understand and interact with, such as a calculation, the eligibility criteria for a benefit [...] or automated financial reporting obligations for compliance.”⁴⁹ Machine consumable legislation is considered a key component in the digital transformation of government.

⁴⁵ For more details see Service Innovation Lab (n.d.-b) *Better Rules and Legislation as Code*; Webster (2021).

⁴⁶ Department of Internal Affairs (2020), p. 15.

⁴⁷ See Service Innovation Lab. (n.d.-a). *About the Service Innovation Lab*.

⁴⁸ The two pieces of legislation concerned the Rates Rebates Act and the Holidays Act, see New Zealand Digital Government (2022).

⁴⁹ Service Innovation Lab (2018), p. 3.

Table 2 Key findings on machine consumable legislation (2018)

1	It is difficult to produce machine consumable rules if the policy and legislation has not been developed with this output in mind.
2	An effective way of developing such policy and legislation is for multidisciplinary teams of policy analysts, legislative drafters, service designers and software developers to co-design the policy and legislation, taking a user-centric approach that focuses on how the service could most effectively be delivered.
3	Co-designing rules with policy and service design increases the chances of the policy being implemented effectively and as intended.
4	Machine consumable legislation that is co-developed (i) enables legislation, business rules, and service delivery software to be developed in parallel, ensuring consistency of application, and (ii) increases the opportunities to automate and integrate service delivery.
5	Common frameworks, reference points and data points (like concept and decision models and ontologies) will assist multi-disciplinary teams to co-design policy and legislation.
6	Not all legislation is suitable for machine consumption, but a multi-disciplinary approach will assist in making better rules.

More notably, the motive behind investigating the use and design of machine-consumable legislation is two-fold: one driving force is the inaccessibility of traditional law with its complex jargon preventing citizens and small businesses from fully understanding the legal landscape including its rights and obligations. Another important factor is the issue of efficiency, which is affected by the growing incongruence between traditional law and digitalization developments. As such, most law does not seem to be as effective in fulfilling and realizing its policy intent because it was not designed for the digital environment. This means that interests of citizens and effectiveness are two fundamental aspects motivating the exploration of digital legislation.⁵⁰ Thus, important values that informed the transition to machine-consumable legislation centered around effectiveness and efficiency, as well as guaranteeing an open and transparent government. The move towards machine-consumable legislation took an incremental approach, as the 2018 initiative was followed by a pilot on another piece of legislation in 2019, the Accident Compensation Act 2001. In its Strategy for a Digital Public Service of 2020, the Aotearoa government has embraced this approach by expressing its commitment to adopting digital practices and tools for policy development, particularly investment in “machine-consumable legislation”.⁵¹

From the Service Innovation Lab pilots, six key findings were derived (with several examples thereof mentioned below) (Table 2):

In the case of Aotearoa, it is equally noticeable that a shift towards machine-consumable legislation comes with the consequence that the legislative process needs to be more aware of the digital demands. For example, key finding 1 implicitly hints to the fact that producing machine-consumable legislation without setting this outcome as a clear goal from the beginning is very challenging. This means that

⁵⁰ Ferlazzo (n.d.); Service Innovation Lab (2018), p. 3.

⁵¹ Department of Internal Affairs (2020), p. 26.

the legislative process needs to comply with certain digital requirements and constraints from the very start to achieve a successful synthesis of the legal and digital sphere. This understanding is also implicit in most other key findings produced by the Service Innovation Lab where emphasis is placed on collaboration, co-design, or co-development. Again, the insistence on achieving effective digital transformation of legislation only through collaboration among multi-disciplinary teams implies that the design of legislation is also dependent on the limitations and requirements imposed by the digital environment. In designing legislation that can be applied in the digital environment, it is not sufficient anymore to have lawyers and policymakers involved in the process. It requires additional expertise to understand the constraints and possibilities of the digital realm and to design the law in a way that makes sense to a computer.

More concrete and practical ways in which legislative drafting could produce machine-consumable legislation were specified in the report on the “Better rules, Better outcomes” initiative. For example, legislative content must be adapted to ensure that services can be delivered digitally. This relates to the framing of legislation by paying more attention to the outcome to be achieved rather than the process by which it is achieved, avoiding requirements that cannot be fulfilled digitally, or being overall more technology agnostic in the drafting. In this context, technology agnostic means to defer from being overly specific in means of communication to avoid legislation to become quickly outdated due to a reference to a specific medium of communication that is not being used anymore, such as referring to a fax. Instead, the idea is to be more general when referring to technology to anticipate technological innovation and developments.⁵² Other considerations that reflect the need for legislation to satisfy digital demands is the requirement to mirror the coding structure in the drafting of the legislative document. In other words, the legislative design should to some extent mimic the structure of how it would be encoded into software by way of ordering provisions more according to flow models that reflect hierarchical orders of obligations, or to place non-binary or ambiguous terms further down the decision tree.⁵³ As such, there is a clear recognition that legislative processes will have to orient themselves more on digital requirements to produce effective outcomes in the digital sphere.

4.3 *The Netherlands*

Late 2022, the Dutch government adopted its Value-driven Digitalization Work Agenda (*Waardegedreven digitaliseren*).⁵⁴ This agenda covers many different aspects of digitalization of government (and beyond), but legislative policy is not a

⁵² Service Innovation Lab (2018), pp. 27–28.

⁵³ Service Innovation Lab (2018), p. 28.

⁵⁴ See Ministerie van Algemene Zaken (2022).

central theme here. However, the Council of State, which is the advisory body of the Dutch government on legislative proposals, had already formulated several recommendations on digitalization and legislation in its publication ‘Digitalization. Legislation and Administrative Litigation’.⁵⁵ In response to these recommendations,⁵⁶ the Dutch government has tried to align these recommendations on legislation (which had been a ‘source of inspiration’) with its own work agenda by highlighting several ongoing pilots that reflect the different recommendations of the Council of State, such as the initiative “Wendbare wetsuitvoering” (flexible implementation of legislation) within the Tax Authority.⁵⁷ The Dutch government also hints at the development of a general instrument on ‘implementation-oriented legislation’ by the Ministry of Internal Affairs and Kingdom Relations, which will assist in developing so-called “digitally workable legislation”. As this general instrument has not been finalized yet, we will focus on the general recommendations on legislation made by the Council of State.

In general, the Council of State observes that Dutch governmental bodies seem to conceive the digital transition as an organic development that is progressing more and more and which generates stronger dependencies on digital technologies.⁵⁸ Especially since many public entities are implementing digital solutions independently, there is a considerable disparity regarding the different levels of digitalization among government bodies.⁵⁹ Consequently, the Council of State puts emphasis on two notions in regard to digital legislation. On the one hand, it emphasizes the need for legal compliance with existing laws, legal standards, and principles such as legal certainty, transparency, democratic legitimacy, and overall unity of legislation.⁶⁰ On the other hand, it emphasizes that the ICT domain should be involved in the legislative process from start to finish to nurture and strengthen the collaboration between policy officers, lawyers and IT experts. In this regard, it is recognized that the legislative process must accommodate these new interrelations. Whereas law-making usually follows a traditional linear approach (‘waterfall method’), more iterative and interdisciplinary processes, such as scrum methods, are necessary to strengthen collaboration between legislative drafters and ICT professionals.⁶¹

In sum, the Council of State formulates six legislative recommendations (Table 3):

Within these six recommendations, evidence can be found for ways to make legislation digitally compliant (or ‘digitally workable’). For example, Recommendation 2 suggests that considerations of digitally implementing regulations should be integrated into the legislative process as early as possible. This is a

⁵⁵ See Council of State (Raad van State) (2021).

⁵⁶ See Parliament (2022).

⁵⁷ This initiative has resulted in Ausems et al. (2021); *Wendbare wetsuitvoering* (n.d.).

⁵⁸ See Council of State (Raad van State) (2018).

⁵⁹ Council of State (Raad van State) (2018), p. 17.

⁶⁰ Council of State (Raad van State) (2018), pp. 20–21; (2021), p. 12.

⁶¹ Council of State (Raad van State) (2018), p. 22.

Table 3 Six recommendations on digitalization and legislation

Recommendation 1	Ensure a human-oriented approach when using ICT.
Recommendation 2	Integrate digital technology in the legislative design at the earliest possible stage.
Recommendation 3	Consider when technology needs to be regulated in legislation itself and when delegation is acceptable.
Recommendation 4	Be transparent, consistent and verifiable in the automated implementation of legislation.
Recommendation 5	Ensure a good balance between automated and non-automated implementation of legislation.
Recommendation 6	Pay sufficient attention in advance to the conditions and rules for the support and implementation of government ICT by private parties

good illustration of the bi-directional interaction between the legal and digital sphere. First, this recommendation highlights the benefit in considering from the outset whether ICT allows for better implementation and to assess the limits and requirements that the digital sphere poses on legislation. At the same time, however, it also recognizes that certain legal requirements, such as compliance to data protection or human rights, ought to be upheld when translating law into code.

Interestingly, Recommendation 4 indirectly implies that compliance of legislation to software requirements and infrastructure can be as decisive in the legal effects as the substance of the law. In other words, the call for transparency, consistency, and verifiability in the automated implementation of legislation can be seen as a recognition of the fact that the conversion of legislative texts into software as well as the way software executes legislative rules can create significant legal consequences for all parties involved. This makes it more important to guarantee transparency or verifiability in order to be able to trace back the ways in which digital compliance of legislation may effectuate its enforcement.

Recommendations 3 and 5 urge the legislator to make explicit choices about incorporating technology in the implementation of legislation. Recommendation 5 clarifies that it is also a task of the legislature to weigh the benefits of digital implementation of legislation against the disadvantages thereof. In addition, recommendation 3 highlights that parliament, which is representing the people, should be involved in questions whether and how a certain technology should be applied.

Finally, Recommendation 6 draws attention to the (im)balance between public and private actors in implementing legislation. However, this recommendation also extends to the legislative process itself if digital constraints to legislation are prescribed by private actors who are owners or developers of the technology which is necessary for the automated execution of legislation. Thus, the shift towards digital legislation does not only call for a rearrangement of responsibilities within public institutions (between legislative lawyers and software engineers working together), but also between public institutions and private actors.

5 Comparison

5.1 *Different Concepts: What's in a Name?*

It becomes immediately apparent from the descriptions in the previous sections that these three countries follow different strategies in terms of the extent to which law-making should mold itself to digital demands. Denmark is taking quite a radical *top-down* approach by obliging itself to make all legislation “digital-ready” by default based on seven principles. ‘Digital-ready legislation’ takes the form of a comply-or-explain approach, where “digital-ready” means that any obstacles that may stand between law and translating certain complexities into code will have to be overcome. The approach of Aotearoa can better be characterized as *bottom-up*. Taking some concrete pieces of legislation to experiment with, the first contours of “machine consumable legislation” can be drawn. While machine-consumable legislation clearly goes beyond machine-readable legislation, the Aotearoa approach makes clear that not all legislation is suitable for machine consumption. Compared to the Danish top-down strategy, the Dutch approach can be categorized as more pragmatic and less digitally oriented. Similar to Aotearoa, the Dutch approach does not automatically assume that digital-ready legislation is always the preferable option. Rather, a balancing approach is taken (see Recommendation 5) that assesses the compatibility of the legislation with possibilities of automation. Interestingly, the recommendations of the Dutch Council of State are phrased as recommendations on ‘digitalization and legislation’, thereby leaving both entities seemingly unrelated by keeping them separated. Instead of actively pushing for digital-ready legislation, the Council of State advises the Dutch government to think in more general terms on its legislative policy regarding digitalization beyond the concrete pilots the Dutch government is already running. The Dutch approach could therefore best be understood as *middle-out*.

One pressing question therefore must be answered: can the terms “digital-ready” legislation, “machine consumable” legislation, and “digital-workable” legislation be used synonymously? Although all three frontrunning jurisdictions aim for similar objectives in terms of reconciling law-making and legislation with digital demands, there are subtle variations in how the concepts are used and understood. The Danish digital-ready legislation can be seen as the most dedicated approach to blending the legal sphere with the digital one by doing everything possible to mold legislative instruments in a way that makes them operationalizable through code. At the same time, Aotearoa’s machine-consumable legislation is taking a balanced approach by considering where the effectiveness of legislation and public service delivery can benefit from digitalization. In contrast, the Dutch notion of digital-workable legislation appears more conservative in comparison by displaying a degree of caution regarding adapting the legislative process too much to digital constraints. Although these differences are merely nuances in the development of merging legislation and digitalization, it is still important to point these out and to

Table 4 Digital-ready legislation compared

	Denmark	Aotearoa (New Zealand)	The Netherlands
Initiative	Parliament	Government	Council of State
Approach	Top-down	Bottom-up	Middle-out
Concept	Digital-ready legislation	Machine-consumable legislation	Digital-workable legislation

demonstrate how they affect the strategic and procedural decisions taken by these three jurisdictions (Table 4).

5.2 Who Has Supremacy? Democracy or Technocracy

Returning to our previous discussion on the tension between legislative supremacy and the involvement of technocrats in the law-making process, some further reflections will be provided. Especially with a rise of legislative instruments being integrated into the digital realm, the tension between traditional legislative processes backed by the notion of democracy and increasingly necessary input of technical experts (often from the private sector) in terms of translating law into software becomes more prevalent. In this context, it is therefore interesting to ask: should all legislation be digital-ready or machine-consumable? And what are the consequences for the democratic sphere?

Apparently, Denmark seems to answer this question in the affirmative, as all legislation must be digital-ready (at least by default), and therefore, must comply with the requirements and limitations of the digital realm.⁶² Although Denmark is arguably the strongest opponent of digital-ready legislation of the three frontrunners that were introduced, this does not automatically result in strict technocratic supremacy. While there is clearly a push for making legislation more digitally compliant by relying on objective criteria and notions, such as uniformity and coherence, there is also the acknowledgment that certain situations require a degree of discretion which cannot be formalized into code.⁶³ This means that in areas that the legislature deems essential in being decided on a case-by-case basis, automation will be precluded, and human discretion should be preserved. Nevertheless, the aim of making all legislation digital-ready has important repercussions that shift a considerable share of the decision-making power to technical experts and thus in favor of technocracy, as many legislative decisions will have to be assessed in light of digital compliance.

In contrast, Aotearoa recognizes that there is a distinction between legislation where it makes sense to convert it to into a machine consumable version and legislation where that is not the case. Subsequently, the pilots report some legal

⁶² Scupola and Mergel (2022), p. 5.

⁶³ Agency for Digital Government (2018a), p. 11.

characteristics which can give a hint at whether digitalization is beneficial or not, including *inter alia* when legislation involves some form of calculation, repeated processes that are prescribed, or objective criteria as determining factors.⁶⁴ Generally speaking, Aotearoa appears to take a balanced approach in which digitalization of legislative instruments and public service provision is seen as a way forward but not as a silver bullet. In its Digital Strategy, the government clearly states that not everything has to be digital, and that trust remains an important aspect in its pursuit of creating a digital government.⁶⁵ Moreover, even if public services are provided digitally, emphasis is put on making these digital services and decision-making processes compliant with legal obligations, including human rights.⁶⁶ As such, Aotearoa is taking another, compared with Denmark more balanced approach which seeks to optimize democratic mechanisms with the help of technocratic support while carefully considering when machine-consumable legislation can be beneficial.

The Netherlands, in comparison, can be described as more conservative and pragmatic in the way that it assesses the potential of digitally transforming law. In its recommendations, the Council of State somewhat sets the tone as to how digitalization is perceived: “Lawyers and policymakers generally assume that a digitalized implementation offers room for flexibility, variety and infinite possibilities. However, it is exactly the opposite: digitalized implementation required precision, standardization and centralization, because this is the only way to gain the benefits of digitalization.”⁶⁷ Accordingly, the limitations of the digitalization of law are clearly highlighted. Moreover, it seems as if the Dutch strategy is more focused on redefining the legislative process rather than the form of legislation to accommodate the digital transition. While avenues of translating law into code are explored in the Dutch case, the emphasis remains on multi-level stakeholder involvement and comprehensive translation of law into code ensuring that these processes can be retraced. This is intended to enhance both transparency and trust as changes resulting from digitalizing legislation are accounted for to guarantee legal certainty.⁶⁸

In light of these observations, it can be said that both the technocratic and democratic spheres coexist in all three cases presented. Whereas it is true that the shift towards digital-ready or machine-consumable legislation requires more input from the technocratic sphere, this does not negate the notion of legislative supremacy. All three jurisdictions recognize the need for better collaboration and more iterative legislative processes to adapt law-making to the digital era. This may also contribute to preventing asynchronous evolutions of technology as multi-stakeholder involvement will become increasingly necessary in the legislative sphere. Moreover, it is also worth pointing out that efforts to make legislation more accessible by designing it more concisely and comprehensively to tailor it to technical requirements can also

⁶⁴ Service Innovation Lab (2018), p. 27.

⁶⁵ New Zealand Government (2022), pp. 17, 23.

⁶⁶ See New Zealand Government (2019).

⁶⁷ Translated from Dutch, see Council of State (Raad van State) (2018), p. 21.

⁶⁸ Council of State (Raad van State) (2021), p. 18 and see also *Wendbare wetsuitvoering* (n.d.).

contribute to strengthening the democratic process. The tension between and discussion around the democratic and technocratic sphere is in no way just black and white. Whereas Denmark, Aotearoa and the Netherlands apply different balancing exercises when it comes to legislative supremacy and technocratic involvement, there is an overall recognition that some areas are more suitable for the implementation of machine-consumable legislation than others. Consequently, even though it is important to be cognizant of the tension between democracy and technocracy, there are various ways in which they can coexist.⁶⁹

5.3 Safeguarding Virtues of the Legislative Process in Digital-Ready Legislation

Now that we have discussed three frontrunning jurisdictions in detail and examined their different approaches to shifting towards digital-ready legislation, this last section will offer some reflections on what this transition means for guaranteeing the key virtues that are usually associated with legislation, at least in Western liberal democracies. By comparing these three countries and their approaches, some lessons can be learnt which will be synthesized hereafter.

1) Democracy

To recap, the virtue of democracy entails that political power is exercised by the people in an equal manner and it is closely related to the principle of procedural equality.⁷⁰ Shifting the legislative process towards digital-ready legislation has no direct consequence for procedural equality: every citizen remains to have an equal vote in the process. However, certain safeguards need to be put in place to ensure that political power can be effectively exercised by the people and their representatives. In other words, it needs to be guaranteed that lawmakers are actively involved in the legislative design by collaborating with technical experts instead of delegating their tasks. Efforts to secure effective political power can be seen in all examples presented, as all three jurisdictions put strong emphasis on collaboration instead of isolated processes that separate legislative drafters from technical experts. However, helpful in this regard could be to strengthen overall digital literacy to counteract dependencies on technocrats and their expertise.⁷¹

2) Deliberation and compromise

⁶⁹ Gilley (2017), p. 18.

⁷⁰ Beetham (1999), pp. 1–29; Daly (2017), pp. 501–531.

⁷¹ For example required under Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act) (2024), art. 9.

Deliberation and compromise are a second virtue attributed to the legislative process in general and the political arena in particular. However, the shift towards incorporating ‘digital demands’ in the process of drafting a legislative proposal, can somehow freeze the legislative proposal when it is submitted to parliament. For example, if a legislative proposal has been developed by lawyers and technical experts such as to be machine-consumable, what room is left for politicians to amend the legislative proposal without undermining a successful digital implementation of the legislation? At the same time, the analyzed jurisdictions (Denmark and the Netherlands in particular) stress the need for discussing the digital implications of legislation in parliament, also because such an evaluation of legislation is an important sanity check for the quality of the legislation in general.

3) Purposive generality

Digital-ready or machine-consumable legislation seems to be perfectly in line with the virtue of purposive generality. The Danish concept of digital-ready legislation does not only stress the need for using clear and consistent wording, but also argues that exceptions should be preserved to the extent that it is necessary to ensure that the civil rights of citizens are protected. It suggests that instead of additional exceptions and special schemes, a basic revision of the legislation should be made to make it generally simpler and clearer. Equally, the Aotearoa pilot clarifies that the power of digital legislation is not to simplify and thus remove a nuanced response to complexity, but rather to clarify and to make the implementation more streamlined and consistent. The purposive generality of legislation even transcends a particular piece of legislation: with the emphasis on using concepts in legislation that are generally applicable, especially since digital systems are usually not developed for one single piece of legislation, the purposive generality of legislation is further strengthened.

4) Transparency and accountability

It is interesting to note that the emphasis on digital-ready or machine-consumable legislation has not been contrasted with human-consumable legislation in the three jurisdictions concerned. Instead, the dominating view is that digital-ready legislation will also serve citizens in understanding legislation, as legal provisions need to be drawn in a simple and consistent manner, whereas they need to be grouped in such a way that machinable consumption is facilitated. Thus, digital-ready legislation might contribute to transparency as far as the outcome (form) of legislation is concerned. At the same time, there are concerns about the transparency of the law-making process. With regard to the Danish concept of digital-ready legislation, for example, it is observed that the complexity of the law-making processes and the range of actors involved in digital-ready legislation seems to be a central reason why important processes and decisions are immersed in darkness.⁷² In more general terms, the emphasis on digital-ready legislation in the drafting process, seems to put

⁷² Plesner and Justesen (2022), p. 158.

more emphasis on the “pre-legislative” stage before the legislative proposal is submitted to and discussed in parliament. Thus, more than being a lawmaker itself, parliament is urged to control those who drafted the legislative proposal.

6 Conclusion

Can AI write laws? While parliaments might not immediately fear that AI will make their legislative work superfluous, the shift towards digital-ready or machine-consumable legislation in three frontrunner jurisdictions (Denmark, Aotearoa (New Zealand), and the Netherlands) undoubtedly shows the potential impact on the legislative process. Even though it would be too far-reaching to argue that legislation is subject to the ‘rule of tech’, technology is gaining ground in the legislative process: whereas technology used to follow legislation (becoming important in the implementation stage), they do now increasingly co-exist in the stage of adopting legislation.

This co-existence of technocracy (technology) and democracy (legislation) raises the question whether the virtues usually ascribed to legislation are still upheld in the case of digital-ready or machine-consumable legislation. This chapter concludes that the effects of digital-ready legislation are not one-directional. On the one hand, digital-ready legislation strengthens the virtues usually associated with the outcome (or form) of legislation, particularly its purposive generality and transparency of the legislative output. On the other hand, digital-ready legislation challenges virtues that are associated with the process of enacting legislation, such as democracy, deliberation, and transparency of the legislative process. The increased attention for collaboration in multidisciplinary teams between lawyers, policymakers, and ICT professionals when drafting legislative proposals to adopt legislation that satisfies digital demands ‘by design’, might require a different attitude of the formal lawmaker (parliament) when evaluating legislative proposals, including a willingness to discuss challenges of digitalization in the implementation of legislation explicitly. Only by transforming the legislative process, technocracy will not prevail over democracy by making digital constraints imperative in the legislative process. Instead, technocracy and democracy could co-exist and work together.

References

- Agency for Digital Government (2018a) Guidance on digital-ready legislation. https://en.digst.dk/media/20206/en_guidance-regarding-digital-ready-legislation-2018.pdf
- Agency for Digital Government (2018b, January 16) Agreement on digital-ready legislation. Agency for Digital Government. https://en.digst.dk/media/20205/en_political-agreement-regarding-digital-ready-legislation.pdf
- Agency for Digital Government (n.d.) About the Agency for Digital Government. Retrieved November 25, 2023, from <https://en.digst.dk/about-us/>

- Allan TRS, Allan TRS (2001) *Constitutional justice: a liberal theory of the rule of law*. Oxford University Press
- Ausems A, Bulles J, Lokin M (2021) Wetsanalyse - Voor een werkbare uitvoering van wetgeving met ICT. Boom. https://www.boom.nl/juridisch/100-11574_Wetsanalyse
- Beetham D (1999) *Democracy and human rights* (Eerste editie). Polity Press
- Boer A, Sileno G (2013) AGILE: a methodology for advanced governance of information services through legal engineering [Technical Report Leibniz Center for Law]. Leibniz Center for Law, p 255. <http://www.leibnizcenter.org/papers/agile-final.2013.pdf>
- Council of State (Raad van State) (2018) Ongevraagd advies over de effecten van de digitalisering voor de rechtsstatelijke verhoudingen (Recommendations W04.18.0230/I). Council of State (Raad van State)
- Council of State (Raad van State) (2021) Digitalisering - Wetgeving en bestuursrechtspraak. Council of State (Raad van State), p 134
- Daly E (2017) Legislative form as a justification for legislative supremacy. *Jurisprudence* 8(3):501–531. <https://doi.org/10.1080/20403313.2017.1296621>
- Daly E (2020) Transparency as a justification for legislative supremacy. *Crit Rev Int Soc Polit Philos* 23(7):807–830. <https://doi.org/10.1080/13698230.2018.1497247>
- De Gregorio G (2023) The normative power of artificial intelligence digital constitution: on the transformative potential of societal constitutionalism. *Indiana J Global Leg Stud* 30(2):55–80
- Department of Internal Affairs (2020) Strategy for a Digital Public Service, p 36
- Digital Nations (2020) Digital Nations - Leading Digital Governments. Digital Nations. <https://www.leadingdigitalgovs.org>
- Diver L (2024) Using design patterns to build and maintain the rule of law. *Digit Soc* 3(2):30. <https://doi.org/10.1007/s44206-024-00115-0>
- Epifanova TV, Vovchenko NG, Toporov DA, Pozdnyshov AN (2020) Development of legal education and machine-readable law in the conditions of economy digitization. In: Popkova EG, Sergi BS (eds) *Digital economy: complexity and variety vs. rationality*. Springer International Publishing, pp 971–979. https://doi.org/10.1007/978-3-030-29586-8_110
- Fenwick M, Kaal WA, Vermeulen EPM (2017) Regulation tomorrow: what happens when technology is faster than the law? *Am Univ Bus Law Rev* 6(3):561–594. <https://doi.org/10.2139/ssrn.2834531>
- Ferlazzo E (n.d.) New Zealand: Strengthening Compliance Through Legislation-as-Code. Ministère de l'Économie des Finances et de la Souveraineté Industrielle et Numérique. Retrieved November 28, 2023, from https://www.economie.gouv.fr/igpde-editions-publications/responsive_public_management_n7
- Francart T, Dann J, Pappalardo R, Malagon C, Pellegrino M (2019) The European Legislation Identifier. In: *Knowledge of the law in the big data age*, vol 317. IOS Press, pp 137–148. <https://doi.org/10.3233/FAIA190016>
- Gaudet LM, Marchant GE (2011) Administrative law tools for more adaptive and responsive regulation. In: Marchant GE, Allenby BR, Herkert JR (eds) *The growing gap between emerging technologies and legal-ethical oversight - the pacing problem*, vol 7. Springer, pp 167–182
- Gilley B (2017) Technocracy and democracy as spheres of justice in public policy. *Policy Sci* 50(1):9–22. <https://doi.org/10.1007/s11077-016-9260-2>
- Government of the Netherlands (2022) Value-Driven Digitalisation Work Agenda. Ministry of the Interior and Kingdom Relations, p 54
- Hogan M (2015, January 26) Law by Design. *Law By Design*. <https://lawbydesign.co/legal-design/>
- Kaal W (2016, September 22) What happens when technology is faster than the law? | CLS Blue Sky Blog. The Columbia Law School Blue Sky Blog. <https://clsbluesky.law.columbia.edu/2016/09/22/what-happens-when-technology-is-faster-than-the-law/>
- Leveson NG (2012) *Engineering a safer world: systems thinking applied to safety*. MIT Press. <https://ieeexplore.ieee.org/book/6267532>
- Ma M, Wilson B (2021) The legislative recipe: syntax for machine-readable legislation. *Northwest J Technol Intellect Prop* 19(1):107–136

- Manoharan AP, Melitski J, Holzer M (2023) Digital governance: an assessment of performance and best practices. *Public Organ Rev* 23(1):265–283. <https://doi.org/10.1007/s11115-021-00584-8>
- Ministerie van Algemene Zaken (2022) Value-Driven Digitalisation Work Agenda [Rapport]. Ministerie van Algemene Zaken. <https://www.government.nl/documents/reports/2022/11/30/value-driven-digitalisation-work-agenda>
- Mohun J, Roberts A (2020) Cracking the code: rulemaking for humans and machines (OECD Working Paper on Public Governance No. 42, p 109) [OECD Working Paper]. OECD
- New Zealand Digital Government (2022, February 8) Better Rules for Government Discovery Report. <https://www.digital.govt.nz/>
- New Zealand Government (2019) Strategy for a Digital Public Service, p 40. <https://www.digital.govt.nz/digital-government/strategy/strategy-summary/strategy-for-a-digital-public-service/>
- New Zealand Government (2022) The Digital Strategy for Aotearoa - Enabling Aotearoa New Zealand's people, communities, economy, and environment to flourish and prosper in the digital era, p 51. <https://www.digital.govt.nz/assets/Digital-government/Strategy/Digital-Strategy-for-Aotearoa-English-PDF.pdf>
- Office of the Press Secretary (2013, May 9) Executive Order -- Making Open and Machine Readable the New Default for Government Information. The White House. <https://obamawhitehouse.archives.gov/the-press-office/2013/05/09/executive-order-making-open-and-machine-readable-new-default-government>
- Parliament (2022) Parliamentary Document “Kamerstukken II 2022-2023, 29 362, nr. 319.”
- Plesner U, Justesen L (2022) The double darkness of digitalization: shaping digital-ready legislation to reshape the conditions for public-sector digitalization. *Sci Technol Hum Values* 47(1):146–173. <https://doi.org/10.1177/0162243921999715>
- Pünder H (2009) Democratic legitimation of delegated legislation: a comparative view on the American, British and German Law. *Int Comp Law Q* 58(2):353–378
- Rangone N (2023) Artificial intelligence challenging core State functions: a focus on law-making and rule-making. *Revista de Derecho Público: Teoría y Método* 8:95–126. https://doi.org/10.37417/RDP/vol_8_2023_1949
- Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act), EU 2024/1689 144 (2024)
- Scupola A, Mergel I (2022) Co-production in digital transformation of public administration and public value creation: the case of Denmark. *Gov Inf Q* 39(1):1–11. <https://doi.org/10.1016/j.giq.2021.101650>
- Service Innovation Lab (2018) Better Rules for Government Discovery Report. <https://www.digital.govt.nz/dmsdocument/95-better-rules-for-government-discovery-report/html>
- Service Innovation Lab (n.d.-a) About the Service Innovation Lab. Retrieved December 21, 2023, from <https://serviceinnovationlab.github.io/about/>
- Service Innovation Lab (n.d.-b) Better Rules and Legislation as Code. Retrieved December 21, 2023, from <https://serviceinnovationlab.github.io/projects/legislation-as-code/>
- Sidorenko EL, von Arx P (2020) Transformation of law in the context of digitalization: defining the correct priorities. *Digit Law J* 1(1):24–38. <https://doi.org/10.38044/DLJ-2020-1-1-24-38>
- Suzor NP (2019) Lawless: the secret rules that govern our digital lives. Cambridge University Press. <https://doi.org/10.1017/9781108666428>
- Terzis P (2024) Against digital constitutionalism. *Eur Law Open*:1–17. <https://doi.org/10.1017/elo.2024.15>
- Waddington M (2019) Machine-consumable legislation: a legislative Drafter's Perspective - Human v Artificial Intelligence. *Loophole - J Commonwealth Assoc Legis Counsel* 2:21–52

Webster N (2021, December 7) LabPlus: Better Rules for Government Discovery Report. New Zealand Digital Government. <https://www.digital.govt.nz/blog/labplus-better-rules-for-government-discovery-report/>
Wendbare wetsuitvoering (n.d.) Retrieved November 27, 2023, from <https://wendbarewetsuitvoering.pleio.nl/>

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.



The Shaping of AI Policies in Latin America: A Study of International Influence and Local Realities



María Lorena Flórez Rojas

Abstract Several Latin American countries are investing in developing artificial intelligence (AI) frameworks or strategies. Besides the widespread mention of ethical guidelines, one can identify the main regulatory paths that this region has taken about AI Governance. The chapter describes the AI public policies of the region's countries as well as the internal and external stakeholders involved. It uses a qualitative methodology to assess AI policies and strategies in Argentina, Colombia, Chile, Costa Rica and Brazil, describing the regulatory influence of international organizations in the development of their national policies. After this description, the chapter presents a debate about the limitations and challenges about the primary influence of external parties in their national strategies. Drawing on a risk regulation framework that emphasizes monitoring, standard-setting, and behavior modification, the chapter advocates a shift from traditional multistakeholderism to networked governance. This approach promotes transparency, trust, and genuine local engagement, ensuring that AI policies foster both technological innovation and social justice in Latin America. It concludes that the countries gave little space to assess social and cultural challenges within their national AI strategies due to the influence of international values diminishing their own national values and cultural approaches.

1 Introduction

In June 2020, Santa Cruz, California, made history by becoming the first U.S. city to prohibit the use of predictive policing in its municipal law enforcement.¹ The program faced criticism for exacerbating racial inequalities, prompting the city

¹ See Guariglia (2020).

M. L. Flórez Rojas (✉)
Faculty of Law, Groningen University, Groningen, Netherlands
e-mail: m.l.florez.rojas@rug.nl

council's unanimous decision to ban it after a nine-year run. Predictive policing operates as a self-fulfilling prophecy, concentrating police efforts in specific neighbourhoods based on historical data, which, in turn, amplifies reported criminal activity in those areas. This approach tends to disproportionately impact already overpoliced communities, such as those of colour, unhoused individuals, and immigrants, under the guise of scientific legitimacy.²

Meanwhile, various Latin American countries have adopted predictive policing systems. For instance, in Montevideo, Uruguay acquired the *Predpol* policing software, the same software used in California that was criticized for its problematic opacity and potential reinforcement of discriminatory practices.³ Remarkably, within three years, the Ministry of the Interior discontinued *Predpol*, and instead, they decided to use their own developed statistical tools.⁴ In Bogotá, Colombia, a system based on the Mohler model is being tested, constructed on three assumptions: (1) criminality concentrates in specific areas of the city, (2) higher incidence of crime at certain times and days, and (3) crime spreads like a disease from one place to another.⁵ While this model is still a pilot program aimed at efficiently allocating scarce law enforcement resources, it has raised concerns about stigmatization, discrimination, and the criminalization of vulnerable populations.⁶ Another example is Rio de Janeiro, Brazil. In 2016, the Igarapé Institute collaborated with private actors to develop the *CrimeRadar* app, a crime prediction platform assessing crime frequency in the metropolitan region. The platform provides a visual representation of security levels at specific locations and times.⁷ Nevertheless, concerns have arisen about the potential reinforcement of harmful biases when using these tools within the context of inequality in the country. The significance of the flagged incidents is often influenced by what people see in the news.⁸ These cases exemplify the idea that Policy Design is the initial step in creating a new government policy or making changes to an existing one.

Policy Design refers to the process of formulating and developing government policies to address specific societal issues or achieve certain goals.⁹ It involves identifying problems, setting objectives, analysing options, and crafting strategies and

² Shad (2023), pp. 1–12.

³ Castets-Renard (2021), pp. 96–98.

⁴ Ortiz and Iglesias (2018), pp. 26–30. Currently the police personnel rely on DIT reports which they access through the Public Security Management System (SGSP). The SGSP is a broader platform through which all the information related to the police action is managed. The platform interoperates with several information sources of National Institutions and with Security and Police Agencies of MERCOSUR as well as with databases from the Ministry of the Interior. Note that by Resolution No. 5903 of July 20, 2012, the Ministry of the Interior classified as reserved all information and documentation, which involves official police activity.

⁵ Mohler et al. (2011), pp. 100–108.

⁶ Dulce et al. (2018), p. 839.

⁷ Aguirre et al. (2019), pp. 3–6.

⁸ Duarte (2021), p. 203.

⁹ Van Geet et al. (2021), pp. 631–633.

implementation plans.¹⁰ It emphasizes the importance of not rapidly adopting foreign technologies or policies that are not tailor-made for the country's context.¹¹ This process includes selecting technologies, evaluating their effectiveness, and addressing concerns about potential discriminatory practices or reinforcement of biases.

The intersection of predictive policing and the regulatory landscape of AI in Latin America presents a complex and evolving scenario. As evidenced by the adoption of systems like *Predpol* in Uruguay and the Mohler model in Colombia, Latin American countries are actively exploring AI-driven solutions for law enforcement. This trend in predictive policing mirrors the broader AI regulatory landscape in Latin America, as highlighted by initiatives like the Santiago Declaration.¹² The region is moving towards a more proactive stance in AI governance, seeking to develop frameworks that address its unique socioeconomic and cultural challenges. The proposed intergovernmental Council on Artificial Intelligence for Latin America and the Caribbean reflects a commitment to balancing technological advancement with ethical considerations and human rights. As Latin American countries introduce AI-related legislation and work to overcome technological and infrastructural challenges, the regulatory approach to AI, including predictive policing systems, is likely to evolve.

It is intriguing that despite Latin American countries not ranking high in terms of development and technology, they stand out in creating public policies to regulate the field. According to the 2023 Oxford Insights assessment of 181 countries worldwide regarding their readiness to implement AI in public services, Latin America holds the 6th position among nine regions. This ranking is based on the analysis of three key factors: (1) Governmental strategies, (2) Technological development, and (3) Data and Infrastructure.¹³ Additionally, the AI index report for 2023 from The Stanford Institute for Human-Centred Artificial Intelligence (HAI) shows that the development of national strategies and regulatory initiatives is not aligned with the technical development of the region.¹⁴ Therefore, this contribution reflects on the influence of international actors on Latin American policies, overshadowing important regional issues such as cultural, social, and environmental risks.

¹⁰The concept of policy by design has been explored by different authors. See, for instance, Mintrom and Luetjens (2016) and Hagan (2020).

¹¹Capano and Lepori (2024), pp. 56–58.

¹²Cumbre Ministerial y de Altas Autoridades de América Latina y el Caribe (2023), pp. 1–4. The “Santiago Declaration for Promoting Ethical Artificial Intelligence in Latin America and the Caribbean” was recently signed by 20 nations during the High Authorities Forum on AI, organized by the Government of Chile, UNESCO, and CAF. The agreement included ministers and officials responsible for digital and AI policies from Argentina, Brazil, Chile, Colombia, Costa Rica, Cuba, Ecuador, El Salvador, Guatemala, Honduras, Jamaica, Mexico, Peru, Paraguay, the Dominican Republic, Saint Lucia, Saint Vincent and the Grenadines, Suriname, Uruguay, and Venezuela.

¹³Hankins et al. (2023), pp. 15–17.

¹⁴Maslej et al. (2023), pp. 267–269.

This chapter establishes a networked governance approach for AI policy in Latin American countries, building upon the author's previous research titled "*Design Thinking and Ethical Frameworks for Artificial Intelligence: Exploring Multi-Stakeholder Engagement*". The earlier study aimed to enrich the Colombian discourse on authentic participatory design within AI strategy, emphasizing how collaborative approaches can enhance the nation's AI strategy.¹⁵ In alignment with this perspective, the chapter integrates prior analysis by Giacalone on the influence of international actors on policy making for Latin American countries, as discussed in her work titled "*Latin American Foreign Policy Analysis: External Influences and Internal Circumstances*".

Giacalone highlights the development of Latin American thought on International Relations (IR), identifying three key factors. Firstly, geographical location influences issue hierarchy, particularly the interest in the role of the United States (U.S.) due to its proximity and power. This focus stems from the historical and ongoing influence of the U.S. in the region's political, economic, and social affairs.¹⁶ Therefore, Giacalone emphasizes how this geographical reality shapes diplomatic strategies, economic policies, and security concerns across the continent. The proximity to the U.S. has led to both opportunities for cooperation and challenges related to sovereignty and autonomy, making it a central theme in Latin American IR discourse.¹⁷

Secondly, Latin American IR is shaped by the creative appropriation of American and European theoretical approaches, adapting these frameworks to regional realities. Thirdly, the region's insertion into the international sphere has determined common themes in IR studies, such as development, regional integration, and global power dynamics. Latin American policy analysis is thus framed by external theoretical influences and domestic circumstances, resulting in a coexistence of analytical and politicized studies.¹⁸ Thus, the current research analyses the concept of networked governance. It examines the role of international actors as stakeholders and addresses their direct and indirect influence on the development of AI strategies in Latin America. Building on this foundation of Latin American IR thought, this contribution extends the analysis to the realm of AI governance. Specifically, it examines the concept of networked governance in the context of AI policy development. By doing so, it aims to illuminate how the interplay between external

¹⁵ Flórez Rojas (2023), pp. 21–25.

¹⁶ Riggiozzi and Tussie (2011), p. 14. In other words, while there are experiments with heterodox policies in some countries in the region, it may be too early to consider an emerging post-hegemony as a synonym for post-neoliberalism. On the other hand, it is easy to exaggerate the loss of influence of the United States in the region. Although the failure to conclude a free trade area for the Americas (FTAA) was a sign for many of the decline of US power in the region, the latter adapted to the situation by concluding free trade agreements with countries of Central America and the Dominican Republic (DR-CAFTA) as well as trade agreements with Chile, Colombia and Peru.

¹⁷ See also Piccato (2010), p. 171.

¹⁸ Giacalone (2012), pp. 340–345.

influences and internal circumstances shapes the region's approach to AI governance, reflecting the broader patterns observed in Latin American IR studies.

The structure of this contribution is based on the proposed framework of risk regulation made by Hood et al., which views control systems as having three essential components: information-gathering or monitoring, standard-setting, and behaviour modification.¹⁹ These components are necessary for a system to be under control. However, linking these components together can be challenging in human organizations due to underlaps, conflicts, and communication failures. It is crucial to emphasize that Latin American countries, especially in the context of legislative and policy considerations, should not be perceived as a uniform entity. The countries within the region exhibit substantial variations in terms of available resources, economic situations, and relative geopolitical power. These factors significantly shape their distinct priorities and operational dynamics. Consequently, any recommendations made cannot be broadly generalized. Currently, Latin American countries find themselves at different developmental stages concerning AI regulations. While some nations have already developed comprehensive AI strategies and are actively working on corresponding regulations, others are in the early phases, formulating emerging AI strategies with the possibility of developing action plans. Importantly, there are those within the region that have not yet established a formal strategy for AI. Furthermore, the maturity of democracies and the level of respect for fundamental rights differ across the region. It is within this diverse landscape that nuanced and tailored approaches to AI regulations and governance must be carefully considered.

Having established the main framework, this contribution addresses the following fundamental question: To what extent have international actors influenced the creation and regulation of AI policies in Latin American countries, and what specific roles have they played? To answer this question, the chapter is divided into three sections. Firstly, it describes the current state of public policies, strategies, and legislative proposals related to AI in Latin America (information gathering). However, the effectiveness of these policies is beyond the scope of this chapter. Secondly, it analyses the concept of networked governance and to which extent a new actor has stood out in the development of Latin America (standard-setting). Then, the chapter examines the respective influence of international actors and how this authority might impact national developments (behaviour-modification). Finally, the chapter presents some conclusions and proposals for further research. The research methodology employs a deductive approach, analysing eight national AI strategies during their creation stage, involving participation from international actors. The study relies on legal dogmatics to scrutinize legal doctrines, principles, norms, and concepts related to AI strategy creation. The objective is to determine how international influence may impact their alignment with national values.

¹⁹Hood et al. (2001), p. 24.

2 Information Gathering: Strategies, Policies, and Legislative Proposals for AI in Latin America

Latin America, a collective region in the Americas, is characterized by the predominant use of Romance languages. In a narrow sense, it encompasses Spanish America and Brazil and is comprised of 20 countries, though not all will be analysed within the scope of this research. The integration of AI into Latin America's cultural, political, and socioeconomic landscape presents a multi-layered challenge with both promising opportunities and substantial risks. Between 2020 and 2023, several countries in the region have formulated national AI strategies or policies, intending to provide guidelines for the responsible implementation of this technology. These national strategies outline AI-related priorities and goals, and in some instances, provide a roadmap for achieving them. Although most initiatives declare to be open to respond to ethical concerns in AI, the reality is that they more often respond to market pressures to guarantee legal environments that can foster the adoption of AI products and services. Thus, this section starts with a general analysis of AI governance schemes in Latin America, subsequently examining specific countries determine the international influence in the AI governance strategy of the country.

In October 2023 the Santiago Declaration was signed in Chile by 20 Latin American countries reflecting their commitment to ethics in AI. It is based on UNESCO's Recommendation on the Ethics of Artificial Intelligence. This document was the main outcome of the Summit of Ministers and High-Level Authorities of Latin America and the Caribbean on AI, organized by THE Development Bank of Latin America (CAF), UNESCO, and the Chilean government. The declaration acknowledges the need for proactive involvement from governments to harness the opportunities that AI brings while addressing its risks.

According to the Oxford Index 2023, the region's performance stands out, especially in the Government pillar, with four out of the five regional leaders ranking within the global top 40 in this category.²⁰ Among Latin American countries, seven have either developed or are in the process of creating a national AI strategy—Argentina, Brazil, Chile, Colombia, Mexico, Peru, and Uruguay (see Table 1). Additionally, despite not currently having a national strategy, Costa Rica has taken a significant step by presenting two bills. At the time of writing, the first bill is undergoing legislative processes. It is worth emphasizing that the first bill was drafted by ChatGPT4.²¹ For some, this trend signals an improved emphasis in this region on maintaining competitiveness with global peers, showcasing a determination to keep pace with the swiftly evolving technology and its potential advantages and risks. However, from a policy design perspective, this trend may also indicate unnecessary pressure to impulsively adopt policy documents without careful

²⁰ Hankins et al. (2023), pp. 15–17.

²¹ Bill 23.771 of 2023 on the Regulation of Artificial Intelligence in Costa Rica. See also Forbes (2023). The Press Conference Presentation of the first Bill Prepared with Artificial Intelligence is available in: <https://www.youtube.com/watch?v=YZ-zpNjr2kQ&t=1s>.

Table 1 Mapping strategies and regulatory proposals for AI in Latin America. By the author

Country	National Strategy/Policy	Legislative proposal
Argentina	National Artificial Intelligence Plan (2019)	Bill Modification of National Law 25.467 proposes to add three articles that establish principles and values for the “ethical functioning” of AI systems. Bill about Legal framework for the regulation of the development and use of Artificial Intelligence, 2505-D-2023. Provision 2/2023, DI-2023-2-APN-SSTI#JGM, approving the “Recommendations for Trustworthy Artificial Intelligence”.
Bolivia	No	
Brazil	Brazilian Strategy for Artificial Intelligence (EBIA) (2021)	Bill 5.051/2019: Establishes the principles for the use of Artificial Intelligence in Brazil. Bill 21/2020: Establishes foundations, principles and guidelines for the development and application of artificial intelligence in Brazil; and makes other provisions. Bill 872/2021: Provides for the ethical frameworks and guidelines that underpin the development and use of Artificial Intelligence in Brazil. Bill 2338/23: On the use of Artificial Intelligence.
Chile	National Artificial Intelligence Policy (2021, 2024)	Bill 15869-19/2023: Regulates artificial intelligence systems, robotics and related technologies, in their various fields of application. Bill 15935-07/2023: Amends the Penal Code to punish the misuse of AI. Bill 16.021-07/2023, which amends the Penal Code to incorporate, as an aggravating circumstance of liability, the use of AI in the commission of an offence.
Colombia	Ethical Framework for AI (2021)	Bills 021/2020C; 354/2020C; 253/2022. Those seek to establish the “public policy and guidelines for the development, use and implementation of AI”. Status: Withdrawn/Archived. Bill 200/2023 C: Whereby AI is defined and regulated, limits are set on its development, use and implementation, and other provisions are enacted. Status: Withdrawn/Archived. Bill 059/2023S: Establishing public policy guidelines for the development, use and implementation of AI. Status: Approved in first debate on November 14, 2023. On March 13, 2024, the proposal for second debate was published in Official Gazette 233/2024. Bill 091/2023S: Whereby the duty of information for the responsible use of AI in Colombia. Status: Approved in first debate. Bill 130/2023S: Whereby the harmonization of AI with people’s right to work is created. Status: Approved in first debate. Bill 225/2024S: Whereby article 296 of Law 599 of 2000, Colombian Penal Code, is modified and established as an aggravating factor. Status: Approved in first debate. The motion submitted for second debate was published in Official Gazette 111/2024.

(continued)

Table 1 (continued)

Country	National Strategy/Policy	Legislative proposal
Costa Rica	Signed a letter with UNESCO to implement “Recommendation on Ethics in Artificial Intelligence”, also supported by the Andean Development Cooperation (CAF) (2023)	Bill 23.771/23: Law to regulate Artificial Intelligence: Status: Presented to the Legislative Assembly. Bill 23.919/23: Law for the Responsible Promotion of Artificial Intelligence in Costa Rica: Status: Presented to the Legislative Assembly
Cuba	AI Development Strategy in Cuba (work in progress 2023)	
Dominican Republic	National Artificial Intelligence Strategy (2023)	
Ecuador	No.	Draft Organic Law on the Regulation and Promotion of Artificial Intelligence in Ecuador.
El Salvador	No.	
Guatemala	No.	
Haiti	No.	
Honduras	No.	
Mexico	IA-MX Strategy (2018)	Draft Bill that Regulates the Use of AI. Parliamentary Gazette. Citizen Legislator Initiatives. LXV/2SPR-5-3226/135000. Draft decree issuing the Law on Ethical Regulation of Artificial Intelligence and Robotics.
Nicaragua	No.	
Panama	Exploratory discussions with OCDE.	Draft Bill No. 014/2023: Regulating Artificial Intelligence in the Republic. Presented to the National Assembly by citizen participation.
Paraguay	No.	Draft project on Artificial Intelligence in Paraguay. 2023.
Peru	National Strategy for Artificial Intelligence (ENIA) (2021)	Law 31814/2023: Promotes the use of AI in favour of the economic and social development of the country.
Uruguay	Artificial Intelligence Strategy for digital government (2019)	Bill 1737/2023 Whereby the Systems using Artificial Intelligence.
Venezuela	No.	

guidance of a comprehensive policy-making process. This rush could potentially result in harm to the countries, leading to the expenditure of resources without a conscious design process that could significantly impact society.

2.1 Argentina

In 2018 efforts were made to design a National AI Plan to provide a general framework for various AI-related areas and promote its development.²² The 2019 National AI Plan, coordinated by the former Secretariat of Science and Technology conducted a survey of AI strategies, particularly focusing on the experiences of the U.S., Japan, France, Canada, the UK, and South Korea. This plan aligns with the “Agenda Digital Argentina 2030” (Decree 996/2018) and the “Argentina Innovadora 2030” plan.²³

Key actors, including representatives from academia, private sector, government, and civil society, were initially invited, with around 80 participants. In early 2019, the first discussion on the National AI Plan involved working groups focusing on different strategic axes.²⁴ Subsequently, through a consultative process of nine months, nearly 400 people participated in discussions on 13 strategic pillars. While it resulted from a consultative process and the work of various government bodies, it did not officially become a document.²⁵ As of now, the plan has advisory status, and the new presidential administration has not stated how it plans to proceed with this instrument at the national level.

Although the strategy document currently serves as a “reference document,” analysing the plan’s design process and the involved actors are crucial for this contribution. In general terms, various consulted sectors highlight the participatory nature of the National AI Plan’s development process and the inclusion of actors from diverse sectors. However, concerns have been raised about the lack of transparency regarding the criteria used to *invite actors*, set thematic lines for discussions, and determine which topics would be included in the final document. Ambiguity about the relevance of the participation of different actors has also been emphasized. Finally, on June 2, 2023, the Subsecretariat of Information Technologies adopted the Provision 2/23, aligning with the *UNESCO Recommendation on the ethics of AI*.²⁶ This manual was designed to provide a framework for the technological adoption of AI centered on citizens and their rights, considering its social and strategic aspects, emphasizing the importance of ethical guidelines in shaping the country’s approach to AI.²⁷

²² See Argentine Government (2018).

²³ See Argentine Government (2022).

²⁴ Aguerre and Daniel (2021), pp. 7–10.

²⁵ See Asociación por los Derechos Civiles (ADC) (2020).

²⁶ See objectives and Annex 1 Argentine Government (2023).

²⁷ See Secretary of Public Innovation (2023).

2.2 *Brazil*

In April 2021, the Ministry of Science, Technology, Innovation and Communication (MSTIC) in Brazil disclosed the Brazilian Strategy for Artificial Intelligence (BSAI/EBIA), a plan aimed at guiding the country's development in AI.²⁸ According to national scholars, this strategy highlights several critical points regarding its design and formulation.²⁹ Meanwhile, according to other international organizations such as the OECD, Brazil is asserted to be dedicated to regulating AI and engaging in global cooperation to advocate for trustworthy AI principles, comparing the proposed AI legislation with the developments in other regions and countries such as the European Union (EU) and the U.S.³⁰ One of the significant aspects is the context in which the strategy was developed, which was marked by vague ideas about AI within a populist political regime. In this regard, the political context of the country, coupled with bureaucratic management resistant to societal interaction, created an environment of institutional weakness and hindered effective policy design.

In addition, the design process of the EBIA, between 2019 to 2020, involved three main approaches: a consultancy by UNESCO, benchmarking international experiences, and a public consultation initiated by the MSTIC. However, the *lack of public sharing* of results from these stages raises concerns about transparency. UNESCO played a crucial role in shaping the problems and benchmarks based on international initiatives.³¹ The design dynamics, however, reflected a technocratic perspective with minimal stakeholder involvement in identifying problems and solutions. Although public consultations were conducted using a digital platform, these suggestions from industry, civil society, and other stakeholders were not systematically incorporated into the EBIA document. Finally, it is important to mention the report by the Federal Court of Accounts (TCU) finding that the objectives set out in the Brazilian AI Strategy (EBIA) were not “not specific, measurable, realistic (achievable)” and that as such the “proper implementation” of the country's AI policies are compromised.³²

²⁸ Ministry of Science Technology and Innovation (2021), pp. 1–16.

²⁹ Magrani (2021), p. 6; see also Gaspar and de Mendonça (2021), pp. 4–6; Filgueiras and Junquilho (2023), pp. 7–10.

³⁰ See Shimoda Uechi and Guimarães Moraes (2023). See also Schertel Mendes and Kira (2023) where the authors stated that Brazil must advance discussions on acceptable risk levels for emerging technologies and corresponding mitigation strategies. This process requires public and democratic scrutiny, not just corporate decision-making, and necessitates progress in debates on AI and platform regulation.

³¹ Magrani (2021), pp. 9–10.

³² The TCU decided that it will continue to monitor the level of maturity in artificial intelligence of federal institutions. To this end, the Court of Auditors will develop its own audit reference system to monitor AI. Federal Court of Accounts Judgment 1139/2022. TC 006.662/2021-8.

2.3 *Chile*

Chile published its first version of the National AI Policy in 2021 and has since implemented several initiatives, including the creation of the National Center for Artificial Intelligence (CENIA) and the Millennium Nucleus Futures of Artificial Intelligence Research (FAIR).³³ The associated Action Plan comprises 177 initiatives involving 14 ministries, ensuring effective policy implementation.³⁴ In 2024, Chile conducted an online public consultation focused on the Governance and Ethics axis of the National AI Policy. This consultation aimed to gather citizens' views, perceptions, opinions, and concerns about AI use and development in Chile, ensuring that the policy reflects societal values and concerns.³⁵

The preparation of the first initiative involved two stages of citizen participation: in the first stage, contributions were received from individuals or organizations that wished to contribute their knowledge and experience; in the second stage, a public consultation was held on the first draft of the document.³⁶ It aligns with government-established pillars, addressing factors enabling AI development, the research and application of AI, and ethical, regulatory, social, and economic impacts. The implementation plan, as outlined by the Ministry of Science, Technology, Knowledge and Innovation, includes both short and medium-term measures, encompassing legislative projects, administrative actions, capacity building, and territorial development, with budgets tailored to the nature of each initiative.³⁷ In addition, Chile is set to become the first country in the region to apply the UNESCO Readiness Assessment Methodology (RAM) in updating its national AI policy.

2.4 *Colombia*

The Ethical Framework for AI in Colombia includes the following documents: 1. CONPES 3975 of 2019: National Policy for Digital Transformation and Artificial Intelligence; 2. CONPES 4023 of 2021: Policy for reactivation, repowering, and sustainable and inclusive growth: New commitment for Colombia's future; 3. Ethical Framework for AI in Colombia with 3 iterations of the document; 4. Conceptual Model for the design of Regulatory Sandboxes & Beaches in AI; 5. Task

³³ Ministry of Science Technology Knowledge and Innovation (2021), pp. 12–15. See also Chilean government (2021).

³⁴ Ministry of Science Technology Knowledge and Innovation (2024a), p. 3.

³⁵ The consultation aimed to make visible the citizen contributions to this policy and thus guide future decisions on artificial intelligence in the country, promoting a fair and equitable digital development for all. Ministry of Science Technology Knowledge and Innovation (2024b), p. 9.

³⁶ Ministry of Science Technology Knowledge and Innovation (2021), pp. 10–13.

³⁷ Ministry of Science Technology Knowledge and Innovation (2024c), pp. 12–16.

Force for the development and implementation of AI for Colombia; 6. International Council of AI in Colombia; 7; 8. Mission of AI experts in Colombia; 9. National Digital Strategy 2023-2026; and the Roadmap for the development and implementation of AI 2024.

The development of the National Policy for Digital Transformation and AI was announced on November 8, 2019. The final version of the Ethical Framework was launched on October 12, 2021. It aims to “(...) provide a soft law guide of recommendations and suggestions to Public Entities, to address the formulation and management of projects that include the use of AI.”³⁸ While Colombia has been recognized for developing public policies on AI, diverse civil society organizations have criticized these efforts.³⁹ They argue that while the drafted documents appear promising on paper, the government has failed to implement them effectively. Critics highlight the absence of a genuine strategy for public participation and contextual consideration.

The central issue lies in the lack of involvement of various stakeholders in policy design, resulting in policies that do not adequately address local risks and concerns. This disconnection between policy formulation and practical implementation undermines the potential impact of Colombia’s AI initiatives and raises questions about their relevance to the country’s specific needs and challenges.⁴⁰ Illustrating the international influence on national AI strategies, the Colombian case highlights the challenge of crafting a strategy that aligns with national values, given the predominant tendency to incorporate international perspectives. The use of the word “multistakeholderism” in such cases may not necessarily reflect the authentic national context but, instead, often involves international actors guiding the process.⁴¹ For instance, the mission of AI experts was arranged by people from the Berkman Klein Centre for Internet & Society at Harvard University, The IDB, the Technical University of Munich, the World Bank, the Digital Asia Hub. Notably, the Colombian representation in the mission included people from RappiBank and the

³⁸ Guío et al. (2021), p. 9.

³⁹ Hernández et al. (2022), pp. 22–31. The authors found out that there is a notable lack of mechanisms to engage diverse genders, historically marginalized groups, and rural populations. Insufficient data on representation of priority groups and inadequate accessibility for people with disabilities further exacerbate the issue. While digital platforms are increasingly used for policy consultations, they often fail to address existing technological and literacy gaps. Despite efforts to improve accountability and traceability in consultation processes, the limited scope and lack of commitment in following through on feedback continue to hinder effective participation in AI policy development. See also Mujica Ramírez & Guío Español (2021).

⁴⁰ Camacho (2023), pp. 14–16. The study identifies persistent issues in public participation for AI regulation in Colombia, echoing problems recognized since 2014. Despite some good practices, public entities have not fully or uniformly adopted them. The author suggest that public entities should diagnose participation barriers, internalize the Public Consultation Guide, and implement training and awareness programs. The goal is to transform public consultation from a mere requirement into a co-creative process for addressing state issues and developing beneficial regulatory responses for all stakeholders.

⁴¹ Hernández et al. (2022), p. 24; Camacho (2023), p. 19; Flórez Rojas (2023), pp. 20–23.

Oregon Health & Science University.⁴² It is worth emphasizing that the mission is marked by a substantial presence of foreign experts or Colombians working outside the country.

2.5 *Costa Rica*

The Ministry of Science, Technology, and Telecommunications (MICITT) of Costa Rica, in collaboration with the Inter-American Development Bank (IDB) and with support from the Advisory Council of fAIr LAC Costa Rica, has been developing the National AI Strategy since February 2023. This strategy would serve as a reference document and is expected to provide input for a roadmap that establishes a framework for the ethical, responsible, and secure use of AI in the country.⁴³ This effort is based on the model of the Costa Rican innovation ecosystem through an Advisory Council. This council brings together the private sector (America Free Zone—AFZ, Intel, Microsoft), academia (INCAE), the public sector (CCSS, MEP, MICITT, SINIRUBE, and BID), civil society (CINDE, State of the Nation Program), and the entrepreneurial ecosystem (BID Lab and Innova 10x). In 2023 The MICITT signed a letter of commitment with UNESCO, through which an action plan is agreed for Costa Rica to develop an AI strategy in accordance with UNESCO's "Recommendation on Ethics in Artificial Intelligence", also supported by the CAF.⁴⁴ Both parties, with this letter, are committed to develop a national AI policy, accompany in conducting workshops, identify AI projects in the country.

It is worth mentioning that the recent presentation of a legislative proposal crafted with AI by certain lawmakers in Costa Rica has prompted Minister Bogantes of the MICITT to advocate for prudence in deploying AI tools. During the weekly press conference Bogantes expressed her preference that lawmakers consult MICITT before submitting such proposals, emphasizing the importance of cautious use of AI.⁴⁵ A subsequent clarification from MICITT's Communication Service distinguished the AI-generated document as not a legislative project but rather an overarching strategy. The reference was made in response to a bill presented by Deputy Vanessa Castro and others, which notably was not authored by them or their advisors but rather by the AI tool known as ChatGPT.⁴⁶

⁴² Guío Español (2021), pp. 41–47.

⁴³ Eliana Ulate Brenes et al. (2021), pp. 12–15.

⁴⁴ See UNESCO (2022). See also Álvarez (2023).

⁴⁵ See Code 13 (2024).

⁴⁶ Bill 23.771 of 2023 on the Regulation of Artificial Intelligence in Costa Rica. See also Forbes (2023). The Press Conference Presentation of the first Bill Prepared with Artificial Intelligence is available in: <https://www.youtube.com/watch?v=YZ-zpNjr2kQ&t=1s>.

2.6 *Dominican Republic*

In 2022, the Dominican government launched the National Innovation Policy 2030, envisioning the country's future by addressing three key challenges: Human Talent, Effective Governance, and Innovation Investment. The Effective Governance component aims to create a well-coordinated innovation ecosystem benefiting society, including the development of a National AI Strategy (ENIA) to establish a robust public policy for citizens.⁴⁷ The ENIA was developed with support from two technical consultancies: one from the CAF for methodological guidance and another from the Central American Bank for Economic Integration (CABEI) for hiring an expert consultant in AI policy development.

3 Information Analysis: Influence of International Actors

Various organizations, including RedGEALC, IA Latam, the Organization of American States (OAS), and the Inter-American Development Bank (IADB), are actively contributing to the advancement of AI in Latin America countries. These initiatives are evolving beyond standalone projects, transforming into national plans, policies, or strategies in numerous countries, with some already completed and others in progress. For instance, the United Nations Economic Commission for Latin America and the Caribbean (ECLAC), which is leading efforts to advocate for and propel this transformative movement. Comparably, the OAS is channelling AI to address specific challenges, exemplified by their actions to combat corruption. OAS not only sponsors initiatives like the “Red de gobierno electrónico de América Latina y el Caribe,” a network dedicated to advancing the AI agenda in the region, but also contributes to the broader narrative of AI integration. The landscape also includes non-governmental initiatives such as IA Latam, a network comprising enterprises and researchers, underlining the collaborative and multidimensional nature of the AI drive in the region. Furthermore, even financial institutions like the IADB are actively involved, evidenced by their sponsorship of a report by FairLAC, which examines the state of AI policies across twelve countries in the region.

International collaboration in the development of AI policies can take on two distinct forms: (i) international actors serve as one among many stakeholders, and (ii) international organizations play a leading role in shaping national strategies. The former is exemplified by collaborative efforts such as the World Economic Forum (WEF) and the Organisation for Economic Co-operation and Development (OECD), which contribute to the global discourse on AI but do not dictate national policies. For instance, the WEF has elevated the importance of AI in global agendas, acting more as a catalyst for dialogue than a unilateral policy influencer. The OECD, known for its role in developing science, technology, and innovation policies, serves

⁴⁷ Government of the Dominican Republic (2023), pp. 20–24.

as a forum for expertise and dialogue on AI policy. The OECD AI principles guide policymakers internationally, emphasizing investment in AI research, fostering digital ecosystems, shaping policy environments, building human capacity, and preparing for labour market transformations. However, these organizations operate within the broader context of multistakeholderism, contributing recommendations rather than imposing policies.⁴⁸

On the other hand, international organizations can play a more direct role in leading national strategies, as observed in the case of the OECD. The AI Policy Observatory and the Network of Experts on AI from the OECD have actively contributed to shaping national policies by providing information, data, and multidisciplinary analysis. While this approach ensures a wealth of international expertise, it also raises questions about the extent of sovereign decision-making within national borders.⁴⁹ However, the influence of international organizations may inadvertently shift the focus of national AI strategies toward global agendas, possibly side-lining unique national considerations. Moreover, discussions of global leadership in AI sometimes mimic a competitive space race, potentially diverting attention and resources away from crucial domestic issues. To address this, public policy and governance mechanisms must transform the discourse from a zero-sum game to a positive-sum game, ensuring that the benefits of AI development are distributed globally. Additionally, while competition is a component, cooperation remains a defining characteristic of relations between countries in the AI landscape, emphasizing the importance of collaborative efforts to harness the full potential of AI for the benefit of all.

Nevertheless, civil society organizations in Latin America are currently dealing with a challenge given the international pressure to develop AI strategies. They are resisting UNESCO's guidelines for digital platform regulation, contending that these guidelines inadequately address regional concerns.⁵⁰ The AISur consortium—comprising 11 civil society organizations and academic institutions in Latin America that work to strengthen human rights in the digital environment—sent a letter to UNESCO expressing concerns and disagreements regarding the process of creating guidelines for digital platforms instead of implementing hard law to protect human rights in the region.⁵¹ Although the previous example focuses on a specific aspect of digital transformation—digital platforms—it is analogous to AI strategies, where

⁴⁸ Hernández et al. (2022), p. 32.

⁴⁹ OECD/CAF (2022), pp. 56–60.

⁵⁰ See AISur Consortium (2023). See also Commission of Challenges of the future science technology and innovation (2020), p. 16. Where they mentioned that the primary characteristic of creating a national strategy should be analysing our own realities. Instead of merely copying elements from international initiatives, the focus should be on reflecting our values and strengths.

⁵¹ In the AISur Consortium letter several representatives of civil society recognized the harmful roles tech platforms can play and how government overreach can exacerbate these issues. In our region, rising authoritarianism sees governments leveraging tech platforms to further undemocratic agendas. The letter makes a call for action in different stages: 1. Local over-regulation; 2. Empower authoritarian regulators who lack independence; 3. Incentivize platforms to comply with illegitimate government requests.

many are based on principles and guidelines for companies rather than promoting a clear regulatory landscape.

According to AISur, Derechos Digitales, and Karisma, developing far-reaching proposals risks not enhancing regulation but merely justifying increased regulation, potentially endorsing authoritarian national practices. For instance, Karisma highlights that international influence, such as UNESCO's Recommendations and the EU AI and Data Acts, may shape future regulations in the region.⁵² Similarly, a report by Derechos Digitales on the contribution of comments to national strategies in analysed countries reveals a common pattern: technical discussion tables, often led by CAF, BID, and the Berkman Klein Centre for Internet & Society, were conducted online. This demonstrates that, for some countries, international influence extends beyond preparing a strategy. However, these strategies often lack analysis of their own people and values, focusing more on meeting international standards rather than building a strategy with the citizens.⁵³

In alignment with this, within the Colombian process, participants highlighted the inability to capture local reality in document drafting due to a lack of prior open and diverse dialogue.⁵⁴ A comment on the AI Ethics Framework underscores that the document's preparation involved a single consultant, supervised by government officials, and commissioned by an international financial institution.⁵⁵ In addition, according to Karisma's report the different Colombian documents related to AI, although referring to the country, do not address its realities.⁵⁶ This lack of contextual and local relevance, as well as the absence of evidence tailored to Colombia, was highlighted by several interviewees in the report. For instance, a Colombian health organization stressed the need for documents to go beyond international definitions or initiatives, calling for a collaborative, intersectoral approach to define Colombia's own model.⁵⁷ Similarly, a local robotics centre suggested that Colombian society should choose its principles based on its needs to improve quality of life.⁵⁸ Therefore, this absence of inclusive national dialogue adversely affects social appropriation, legitimacy, and the overall usefulness of both the consultation process and the draft document.

In the Latin American context, it is crucial to identify and understand local realities to formulate feasible and effective recommendations. The previous information gathering reveals the types of risks present in the different Latin American strategies analysed in this chapter. For instance, it shows various social risks, such as insufficient multistakeholder engagement, overreliance on international experts rather

⁵² Policy Network on Artificial Intelligence (2023), p. 16.

⁵³ Hernández et al. (2022), p. 30.

⁵⁴ Camacho (2023), pp. 25–30.

⁵⁵ Administrative Department of the Presidency of the Republic (DAPRE) (2020), cell 145; Flórez Rojas (2023), p. 20.

⁵⁶ Camacho (2023), p. 28.

⁵⁷ Administrative Department of the Presidency of the Republic (DAPRE) (2020), cell 83.

⁵⁸ Administrative Department of the Presidency of the Republic (DAPRE) (2020), cell 175.

than domestic expertise, and the blind adoption of international regulations without considering the country's diversity and technological limitations. Moreover, the data gathering process exposes the interest in regulating emerging areas like AI but fails to address pressing issues such as data protection regimes, consumer protection, and interoperability challenges. Therefore, such a localized approach is essential to mitigate unintended consequences and tailor regulations that effectively address the diverse needs and values of the region.

4 Standard Setting: Networked Governance in Latin American AI Strategies

The process of standard setting is central to regulation, encompassing the establishment of standards, goals, targets, and guidelines. This aspect of regulation is extensively explored in discussions surrounding risk regulation, as it involves determining acceptable levels and distributions of risk. Various approaches to standard setting exist in risk regulation, ranging from technocratic processes based on technical expertise to negotiation-based solutions that represent compromises among stakeholders.⁵⁹ However, based on the information gathered from all the Latin American countries analysed in this chapter, it is apparent that assuming a space labelled as multistakeholderism may not accurately reflect the power dynamics at play. Therefore, a more suitable approach would involve transparency in decision-making processes and adapting regulatory schemes within the framework of network governance.

Networked governance has become a fundamental concept in political science since the 1990s, matching with the emergence of the “network paradigm” in the Internet era.⁶⁰ This concept gained prominence as scholars redirected their attention from theories of “government” to the evolving landscape of “governance.” This approach relies on informal networks to engage citizens and organizations in policy development, implementation, and monitoring. The emergence of the network society, driven by technology, transforms public discourse, decision-making, and policy development.⁶¹ Bang & Esmark suggest that nation-states are evolving into network states embedded in local, regional, and global governance networks. These networks involve *non-state actors* from the private and third sectors as *partners in policy management*, reflecting a shift from conventional democratic polity to governance based on the network paradigm.⁶²

⁵⁹ Hood et al. (2001), p. 30.

⁶⁰ Castells (2000), pp. 693–699. This development means that production is no longer closely tied to the physical space but instead networks exist across local and national borders.

⁶¹ Puppis (2010), pp. 134–149.

⁶² Bang and Esmark (2009), pp. 134–149.

This concept is distinguished from related ones such as “multistakeholderism,” often described as the engagement of two or more classes of actors in common governance with polyarchic authority relations.⁶³ While multistakeholderism involves diverse actors with stakes in a change, network governance goes beyond by emphasizing the negotiation practices among interdependent actors. Both network governance and multistakeholderism include multiple actors in decision-making processes, however, multistakeholderism has faced criticism for not adequately addressing power imbalances among these actors.⁶⁴

Critics argue that it often fails to recognize and address the power relations that contribute to discrepancies in information access. Unlike multistakeholderism, network governance does not assume equal access to information, resources, and support among involved actors. Instead, it recognizes and works with the differences in access, communication methods, and motivations of various stakeholders.⁶⁵ Network governance, in contrast, places a central focus on understanding how policies are continuously shaped and reshaped through the negotiation practices of interdependent actors, offering a more nuanced perspective on power dynamics. Therefore, network governance provides a lens to understand decision-making in a more decentralized and collaborative manner, bridging the conceptual gap between the traditional hierarchical government and the complex public-private neoliberal model.⁶⁶

To make networked governance effective, policymakers need to understand the nature of power in a network structure. Unlike traditional authority, network structures rely on informal exchanges and interpersonal relations.⁶⁷ Leadership in network governance takes on a facilitator or broker role, and trust among members is crucial for mutual benefit. Trust building, risk management, and appropriate indicators are crucial for the success of networked arrangements.⁶⁸ Policymakers should acknowledge that network structures may not deliver traditional results quickly and focus on changing relationships, perceptions, and fostering innovation. Network structures, designed for complex, wicked problems, require a paradigm shift in policymaking, with a recognition of the unique responses they offer.

Across various Latin American countries, the development of national strategies for AI has been significantly shaped by international collaboration and support. Key international actors such as the CAF and institutions like the Berkman Klein Centre for Internet & Society have played pivotal roles in influencing these strategies. For instance, in Colombia, the Berkman Klein Centre contributed to the AI expert mission, indicating a cross-border exchange of insights and expertise. Similarly, UNESCO has wielded substantial influence over most national strategies,

⁶³ Raymond and Denardis (2015), pp. 560–564.

⁶⁴ Caplan (2023), p. 22.

⁶⁵ Sørensen and Torfing (2005), pp. 195–218.

⁶⁶ Eggers (2008), p. 25.

⁶⁷ Wang and Ran (2023), pp. 1190–1195.

⁶⁸ Kapucu and Hu (2020), pp. 49–55.

collaborating closely with some. Additionally, the collaboration with the BID in formulating an AI framework illustrates how international financial institutions contribute to shaping the trajectory of AI policies. The influence, both written and unwritten, of international organizations can accentuate the presence of network governance, rather than a strictly multistakeholder approach, in the formulation of AI strategies in Latin America.

Research on network governance has primarily focused on assessing performance and outcomes within networked settings. The literature predominantly emphasizes measurements such as democracy, deliberation, trust-building, resource utilization, efficiency, effectiveness, and equity, along with factors influencing performance, including network integration, external control, system stability, coordination, and resource availability.⁶⁹ Network governance is gaining prominence due to the valued co-creation involving all stakeholders in public administration decisions. In the context of Latin American AI strategies, one could observe elements of network governance, particularly in the significant role played by international actors in the drafting and implementation processes. However, this influence is often implicit rather than formally acknowledged. Despite the potential power imbalance regarding information and influence, most countries in the region maintain a discourse of multistakeholderism, promoting the idea that all actors have an equal voice in the decision-making process. This apparent disconnect between the reality of international influence and the rhetoric of inclusive governance presents a complex challenge for the development and implementation of AI strategies in Latin America.

Governments of countries analysed in this contribution keep significant autonomy in adopting ideas from international actors. A major constraint on international actors' influence lies in the willingness of national governments to adhere to international norms and standards and to view these organizations as credible sources of policy advice. Even if policy ideas gain traction at the international level, their reception domestically can vary. However, while international actors may not directly shape domestic decision-making, they play a crucial role in diverse practices through reports and conferences, thereby influencing domestic policy actors who advocate for their adoption.⁷⁰ Thus, international actors primarily function as epistemic actors, contributing to policy ideas, while the responsibility for policy implementation remains with domestic political actors and coalitions.

To ensure transparency and enhance rational-legal legitimacy of AI policy development processes, especially in countries where governmental or institutional legitimacy may be challenged, a transparent approach is imperative.⁷¹ Clear and open communication about collaborative efforts involving international organizations, governments, and stakeholders is essential for fostering trust and credibility.

⁶⁹ Wang and Ran (2023), pp. 641–657; Osborne et al. (2021), pp. 650–653.

⁷⁰ White (2020), p. 71.

⁷¹ Consider legitimacy as legality in words of Hoekema and van Manen (2000) as a coherent entirety of convictions that is expressed in the social practice of the legal order and is performed in social institution, concerning the proper construction of society and parts thereof.

However, the perceived lack of legitimacy in some Latin American countries regarding network cooperation in policy design, particularly in the development of AI strategies, poses significant challenges.⁷² Therefore, concerted efforts should be made to enhance the reception of these processes by incorporating transparent mechanisms and actively engaging diverse local stakeholders.⁷³ Additionally, addressing the social and cultural values specific to individual countries is crucial to improving the effectiveness of network governance in the AI policy context. By ensuring inclusivity and transparency in decision-making processes, these challenges related to legitimacy can be mitigated, thus fostering a more stable and receptive environment for network cooperation in AI governance.⁷⁴ This approach not only acknowledges the complexity of AI policy development but also respects the unique socio-political contexts of Latin American countries.

Considering the various direct and indirect influences from international actors in the Latin American context, adopting a networked governance framework for AI policy could create a more inclusive environment. This approach would provide a platform where diverse stakeholders can actively participate in public debates, potentially fostering trust and collaboration among all parties involved in the AI governance process.⁷⁵ Trust facilitates inner network stability, knowledge sharing, and innovation, critical for managing uncertainties inherent in collaboration. By integrating network members, such as the CAF, UNESCO, BID and others, in a collaborative environment, leaders can stimulate communication and connections, promoting collaborative trust. However, challenges persist as issues specific to Latin America, such as the exploitation of click workers and the environmental impact of natural resource extraction for AI enhancement, often go unaddressed, with global north concerns taking precedence. This discrepancy may undermine the legitimacy of AI governance, especially in countries already grappling with issues of trust and transparency. The tension between global AI governance frameworks and local Latin American realities raises important questions about the effectiveness and equity of current approaches.

5 Conclusion

The integration of AI into Latin America's diverse cultural, political, and socio-economic landscape has prompted various countries in the region to formulate national AI strategies or policies between 2020 and 2024. These strategies, while asserting openness to ethical concerns, often reflect a response to market pressures, aiming to create legal environments conducive to the adoption of AI products and services.

⁷² Flórez Rojas and Leal (2020), pp. 206–252.

⁷³ See Botero (2023).

⁷⁴ Caplan (2023), pp. 3465–3466.

⁷⁵ Wegner and Verschoore (2022), pp. 479–499.

While this trend reflects a commitment to staying competitive globally and adapting to rapidly evolving technology, it raises concerns about the potential rush to adopt policies without comprehensive, carefully guided policy-making processes. This rush could lead to drain resources without a conscious design process, potentially impacting society negatively.

Numerous international organizations play significant roles in advocating and steering the transformative movement of AI in Latin America. The region acknowledges the need for a tailored regulatory framework that goes beyond merely detailing rules for AI. The focus is broadened to encompass initiatives promoting digital literacy, technological infrastructure, and research. A significant aspect is the emphasis on public participation and diverse stakeholder involvement, recognizing the transformative potential of AI on individual rights and societal structures.

This research advocates for the implementation of networked governance across all aspects of AI governance mechanisms, emphasizing the importance of trust, legitimacy, learning, power symmetry, and fairness within these frameworks. While designing policies is essential, ensuring effective implementation requires addressing issues of trust among diverse actors, transparency in decision-making processes, and inclusivity in policy formulation. This approach recognizes the interrelation of external and internal influences in shaping governance outcomes, emphasizing the need for collaborative efforts among international organizations, governments, and various stakeholders.

Despite these advancements, persistent challenges need to be addressed. The underrepresentation of Latin American perspectives in global AI and ethics debates raises concerns about the adequacy of frameworks to meet regional needs. The development of AI national strategies is at varying stages across Latin American countries, with some responding to market pressures, potentially prioritizing economic considerations over ethical ones. This indicates that certain AI strategies are shaped more by international pressures than by local realities, potentially leading to unintended consequences for democracy before the formulation of comprehensive AI regulations.

Successful regulations must be rooted in a commitment to human rights, transparency, and accountability, ensuring the responsible, ethical, and inclusive development of AI in the region. Finally, it could be considered a minimum requirement to shift the vision from multistakeholderism to network governance. According to national strategies, the reality suggests that multistakeholderism has not effectively worked across the region due to imbalances of power and information. Therefore, accepting from the co-creation stage that each organization serves as a network point to develop a national strategy could promote legitimacy within the strategies among nationals. This research underscores the need for a more inclusive and context-sensitive approach to AI governance in Latin America.

References

- Administrative Department of the Presidency of the Republic (2020) Matriz de recepción de comentarios para Sandbox y marco ético de IA con sus respuestas. dapre.presidencia.gov.co/dapre/Documents/Matriz-respuesta-comentarios-documentos-IA-201009.zip
- Aguerre C, Daniel ML (December 2021) Public policy reports: Argentina
- Aguirre K, Badran E, Muggah R (July 2019) Future crime: assessing twenty first century crime prediction. Strategic Note 33. https://igarape.org.br/wp-content/uploads/2019/07/2019-07-12-NE_33_Future_Crime.pdf
- AlSur Consortium (28 June 2023) AlSur consortium sends letter to UNESCO on the process of elaboration of guidelines for the regulation of platforms. AlSur. <https://www.alsur.lat/en/blog/alsur-consortium-sends-letter-unesco-process-elaboration-guidelines-regulation-platforms>
- Álvarez M (28 February 2023) Costa Rica será el primer país de Centroamérica en tener una estrategia de Inteligencia Artificial. Unesco News. <https://www.unesco.org/es/articles/costa-rica-sera-el-primer-pais-de-centroamerica-en-tener-una-estrategia-de-inteligencia-artificial>
- Argentine Government (2018) Plan Nacional de Inteligencia Artificial. Observatory of Public Sector Innovation. <https://oecd-opsi.org/wp-content/uploads/2021/02/Argentina-National-AI-Strategy.pdf>
- Argentine Government. (2022). National Plan for science, technology and innovation 2030. <https://www.argentina.gob.ar/ciencia/plan-nacional-cti/plan-cti>
- Argentine Government, Provision 2/2023. Buenos Aires: Office of the head of the cabinet of Ministers Undersecretariat for information technologies, 2023. Accessed: Aug. 01, 2024. [Online]. Available: <https://www.boletinoficial.gob.ar/detalleAviso/primera/287679/20230602>
- Argentine Secretary of Public Innovation. (2023). Recommendations for a Trustworthy Artificial Intelligence. https://www.argentina.gob.ar/sites/default/files/2023/06/recomendaciones_para_una_inteligencia_artificial_fiable.pdf
- Asociación por los Derechos Civiles (ADC) (2020) A human rights approach to the Buenos Aires Artificial Intelligence Plan
- Bang H, Esmark A (2009) Networked governance: why it is different and how it can work. *Adm Theory Praxis* 31(1)
- Botero BC (21 December 2023) Las deficiencias de las políticas de Inteligencia Artificial. El Espectador. <https://www.elspectador.com/opinion/columnistas/carolina-botero-cabrera/las-deficiencias-de-las-politicas-de-inteligencia-artificial/>
- Camacho, L. (2023). Artificial intelligence and its regulation in Colombia: What about citizen participation? <https://web.karisma.org.co/wp-content/uploads/2023/08/informe-inteligencia-artificial-y-su-regulacion-en-Colombia.pdf>
- Capano G, Lepori B (2024) Designing policies that could work: understanding the interaction between policy design spaces and organizational responses in public sector. *Policy Sci* 57(1):53–82. <https://doi.org/10.1007/S11077-024-09521-0>
- Caplan R (2023) Networked platform governance: the construction of the democratic platform. *Int J Commun* 17:3451–3472
- Castells M (2000) Toward a sociology of the network society. *Contemp Sociol* 29(5):693–699. <https://doi.org/10.2307/2655234>
- Castets-Renard C (2021) Human rights and algorithmic impact assessment for predictive policing. In: Micklitz HW, Pollicino O, Reichman A, Simoncini A, Sartor G, de Gregorio G (eds) *Constitutional challenges in the algorithmic society*. Cambridge University Press, pp 93–110
- Chilean government (2021) Política nacional de inteligencia artificial. Chilean Government. https://minciencia.gob.cl/uploads/filer_public/bc/38/bc389daf-4514-4306-867c-760ae7686e2c/documento_politica_ia_digital_.pdf
- Code 13 (10 June 2024) Entrevista: Estrategia de Inteligencia Artificial de Costa Rica. Canal 13. <https://www.youtube.com/watch?v=N09eE2pQq1w&t=303s>
- Duarte DE (2021) The making of crime predictions: sociotechnical assemblages and the controversies of governing future crime. *Surveill Soc* 19(2):199–215

- Dulce M, Ramírez-Amaya S, Riascos Á (2018) Efficient allocation of law enforcement resources using predictive police patrolling. arXiv:1811.12880, 1–5
- Eggers WD (2008) Collaborative governance: a new era of public policy in Australia? In: O’Flynn J, Wanna J (eds) Collaborative governance: a new era of public policy in Australia? ANU Press, pp 1–202
- Filgueiras F, Junquillo TA (2023) The Brazilian (non)perspective on national strategy for artificial intelligence. *Discov Artif Intell* 3(1):1–15
- Flórez Rojas ML (2023) Pensamiento de diseño y marcos éticos para la Inteligencia Artificial: una mirada a la participación de las múltiples partes interesadas. *Desafíos* 35(1):1–31
- Flórez Rojas ML, Leal JV (2020) The impact of artificial intelligence tools: an analysis of the public sector in Colombia. In: Aguerre C (ed) Artificial intelligence in Latin America and the Caribbean: ethics, governance and policies. Article 7 *CeTyS*, pp 206–238
- FORBES (1 June 2023) Costa Rica, presenta ley para regular la IA con ChatGPT4. Forbes Centroamérica. <https://forbescentroamerica.com/2023/06/01/costa-rica-presenta-ley-para-regular-la-ia-con-chatgpt4>
- Gaspar W, de Mendonça YC (May 2021) Artificial intelligence in Brazil still needs a strategy. Cyberbrics. <https://cyberbrics.info/wp-content/uploads/2021/05/EBIA-en-2.pdf>
- Giacalone R (2012) Latin American foreign policy analysis: external influences and internal circumstances. *Foreign Policy Anal* 8:335–353
- Government of the Dominican Republic (2023) Consulta Pública para la Estrategia Nacional de Inteligencia Artificial de la República Dominicana (ENIA) - Foro Comunitario Ciudadano. Government of the Dominican Republic. <https://consultapublica.notion.site/Consulta-P-blica-para-la-Estrategia-Nacional-de-Inteligencia-Artificial-de-la-Rep-blica-Dominicana%2D%2De58e40a872fc4d79a58766fdcf4d14f3>
- Guariglia M (3 September 2020) Technology can’t predict crime, it can only weaponize proximity to policing. EFF. <https://www.eff.org/deeplinks/2020/09/technology-cant-predict-crime-it-can-only-weaponize-proximity-policing>
- Guío Español A (February 2021) Respuestas institucionales para la implementación de la política de inteligencia artificial. Government of Colombia. <https://dapre.presidencia.gov.co/TD/CONSEJO-INTERNACIONAL-INTELIGENCIA-ARTIFICIAL-COLOMBIA.pdf>
- Guío Español A, Tamayo Uribe E, Gómez Ayerbe P (2021) Ethical Framework for Artificial Intelligence. Government of Colombia. <https://minciencias.gov.co/sites/default/files/marco-etico-ia-colombia-2021.pdf>
- Hagan M (2020) Legal design as a thing: a theory of change and a set of methods to craft a human-centered legal system. *Design Issues* 36(3):3–15
- Hankins E, Fuentes P, Livia N, Grau MG, Rahim S (2023) Government AI Readiness Index 2023. Oxford Insights
- Hernández L, Canales MP, de Souza M (2022) Inteligencia Artificial y participación en América Latina: Las estrategias nacionales de IA
- Hoekema AJ, van Manen NF (2000) Typen van legaliteit. Kluwer
- Hood C, Rothstein H, Baldwin R (2001) The government of risk: understanding risk regulation regimes. Oxford University Press
- Kapucu N, Hu Q (2020) Network governance: concepts, theories, and applications. Routledge
- Magrani E (2021) Estrategia Brasileña de Inteligencia Artificial: comentarios sobre la Resolución 4.617/2021 del Ministerio de Ciencia, Tecnología e Innovación (MCTI). Empatía. <https://www.empatia.la/wp-content/uploads/2021/12/ART-DE-OPINION-Brasil-ES-final.pdf>
- Maslej N, Fattorini L, Brynjolfsson E, Etchemendy J, Ligett K, Lyons T, Manyika J, Ngo H, Nibbles JC, Parli V, Shoham Y, Wald R, Clark J, Perrault R (2023) Artificial Intelligence Index Report 2023. The Stanford Institute for Human-Centered Artificial Intelligence (HAI). <https://arxiv.org/pdf/2310.03715>
- Ministry of Science Technology and Innovation (2021) Brazilian Strategy of Artificial Intelligence (EBIA) and Public Policies

- Ministry of Science Technology Knowledge and Innovation (2021) Política Nacional de Inteligencia Artificial. Ministry of Science Technology Knowledge and Innovation. https://www.minciencia.gob.cl/uploads/filer_public/bc/38/bc389daf-4514-4306-867c-760ae7686e2c/documento_politica_ia_digital_.pdf
- Ministry of Science Technology Knowledge and Innovation (2024a) IA Policy Action Plan 2024. Ministry of Science Technology Knowledge and Innovation. https://drive.google.com/file/d/1nNf65hR4Qeq5NqTERdy9rMUwefzOt_lij/view
- Ministry of Science Technology Knowledge and Innovation (2024b) Política nacional de inteligencia artificial. Ministry of Science Technology Knowledge and Innovation. <https://drive.google.com/file/d/1feyZ3GSPyxKXEEvi9aVZA9jS6aHdC-96/view>
- Ministry of Science Technology Knowledge and Innovation (2024c) Política Nacional de Inteligencia Artificial. Ministry of Science Technology Knowledge and Innovation. <https://drive.google.com/file/d/11OLxLp8NyKgpeRFL45X0zStY7SFEJIC/view>
- Mintrom M, Luetjens J (2016) Design thinking in policymaking processes: opportunities and challenges. *Aust J Public Adm* 75(3):391–402
- Mohler GO, Short MB, Brantingham PJ, Schoenberg FP, Tita GE (2011) Self-exciting point process modeling of crime. *J Am Stat Assoc* 106(493):100–108
- Mujica Ramírez MP, Guío Español A (2021) Misión de Expertos en IA de Colombia: Diagnóstico, miembros y objetivos. Government of Colombia. https://consultorsalud.com/wp-content/uploads/2022/07/201021_Mision_de_Expertos_en_IA_de_Colombia.pdf
- OECD/CAF (2022) The strategic and responsible use of Artificial Intelligence in the Public Sector of Latin America and the Caribbean. OECD Public Governance Reviews
- Ortiz FJ, Iglesias C (2018) Algorithms and Artificial Intelligence in Latin America: a study of implementation by governments in Argentina and Uruguay. World Wide Web Foundation
- Osborne SP, Nasi G, Powell M (2021) Beyond co-production: value creation and public services. *Public Adm* 99(4):641–657
- Piccato P (2010) Public sphere in Latin America: a map of the historiography. *Soc Hist* 35(2):165–192
- Puppis M (2010) Media governance: a new concept for the analysis of media policy and regulation. *Commun Cult Crit* 3(2):134–149
- Raymond M, Denardis L (2015) Multistakeholderism: Anatomy of an Inchoate Global Institution. *Int Theory* 7(3):572–616
- Riggirozzi P, Tussie D (2011) The rise of post-hegemonic regionalism in Latin America. In: Riggirozzi P, Tussie D (eds) *The United Nations University Series on Regionalism*. Springer, pp 1–16
- Schertel Mendes L, Kira B (21 December 2023) The road to regulation of artificial intelligence: the Brazilian experience. *Internet Policy Review*. <https://policyreview.info/articles/news/road-regulation-artificial-intelligence-brazilian-experience/1737>
- Shad KB (2023) Artificial intelligence-related anomies and predictive policing: normative (dis)orders in liberal democracies. *AI Soc* 1:1–12
- Shimoda Uechi CA, Guimarães Moraes T (27 July 2023) Brazil's path to responsible AI. OECD. <https://oecd.ai/en/wonk/brazils-path-to-responsible-ai>
- Sipinen M (15 October 2023) Incidencia ante Organismos Internacionales en marco del Policy Network on Artificial Intelligence. Karisma Foundation. <https://web.karisma.org.co/incidencia-ante-organismos-internacionales-en-marco-del-policy-network-on-artificial-intelligence/>
- Sørensen E, Torfing J (2005) The democratic anchorage of governance networks. *Scand Polit Stud* 28(3):195–218. <https://doi.org/10.1111/J.1467-9477.2005.00129.X>
- Ulate Brenes E, Mora E, Pombo C, Rebellón C, Vega N, Lang G, Chaves Martínez P (2021) The importance of establishing a public policy guiding framework for the responsible and ethical use of artificial intelligence and its application in Costa Rica. Banco Interamericano de Desarrollo
- UNESCO (2022) Recommendation on the Ethics of Artificial Intelligence. UNESCO

- van Geet MT, Verweij S, Busscher T, Arts J (2021) The importance of policy design fit for effectiveness: a qualitative comparative analysis of policy integration in regional transport planning. *Policy Sci* 54(3):629–662
- Wang H, Ran B (2023) Network governance and collaborative governance: a thematic analysis on their similarities, differences, and entanglements. *Public Manag Rev* 25(6):1187–1211
- Wegner D, Verschoore J (2022) Network governance in action: functions and practices to foster collaborative environments. *Adm Soc* 54(3):479–499
- White LA (2020) Do international organizations influence domestic policy outcomes in OECD countries? In: Nieuwenhuis R, van Lancker W (eds) *The Palgrave handbook of family policy*. Springer, pp 69–86

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.



Part III
Public Trust and Values: In Technology
We Trust?

Empathy as a Public Value: Overcoming Administrative Vulnerability and Rehumanizing (Digital) Government



Malou Beck and Sofia Ranchordás

Abstract Empathy is a fundamental value in evolutionary biology, human communication, and modern management. This paper argues that empathy, defined here as the ability to relate to others, should also be regarded as a public value and be employed as a response to the vulnerability experienced by many citizens when engaging with digital government, automated public services, and algorithmic decision-making.

Empathy in citizen-government interactions is increasingly important at a time when human interactions have become rare. The remaining human interactions are characterized by decreasing levels of assistance and compassion due to the lack of time, availability, training, and increasing polarization. This is problematic for two main reasons. First, it excludes citizens in vulnerable circumstances, thus ignoring the possibility that life, law, or socioeconomic circumstances can place citizens in situations that require additional understanding and thus empathic approaches. Second, since vulnerability is inherent to the human condition, empathy in the public sector is essential to ensure the adequate pursuit of the public interest.

Drawing on sociolegal and public administration scholarship, we argue that empathy can contribute to good governance as it has the potential to improve government communication, rehumanize government transactions, and contribute to the acceptance of administrative decisions.

Funding: This research was funded by the Dutch Research Council (NWO) in the context of the NWO-Vidi Project ‘Vulnerability in the Digital Administrative State’ led by Sofia Ranchordás.

M. Beck (✉) · S. Ranchordás

Department of Public Law and Governance, Tilburg Law School, Tilburg University,
Tilburg, Netherlands

e-mail: m.e.beck@tilburguniversity.edu; s.h.ranchordas@tilburguniversity.edu

1 Introduction

Law has sought to harmonize two opposing forces since its very inception: the importance of issuing general and abstract laws that do not differentiate between citizens and are thus perceived as fair, just, and equal; and the need to make exceptions for special cases, tailoring the law to specific circumstances. The balance between generalization and particularism is delicate and often difficult to achieve.¹ This challenge is also visible in administrative law, which has long grappled with balancing the imperatives of the public interest against citizens' individual circumstances. On the one hand, administrative law must ensure consistency and the efficient, equal, and fair allocation of public resources while avoiding abuses of power. On the other hand, administrative law should not reduce citizens to data points, numbers, and other quantified elements. Indeed, citizens are not uniform, public policies are intricate and vague, and indeterminate legal terms permit multiple interpretations. This gives rise to the paradox that administrative law should be the same for everyone, but also be open to being, at times, slightly different for everyone, because not everyone is the same all the time. There are moments in life when we may find ourselves unable to make rational decisions, exercise our rights, and engage independently with bureaucracy. Life or circumstances may place us in vulnerable circumstances.

The well-known Dutch Childcare Benefit scandal and the Australian Robodebt², both instances where thousands of citizens were wrongly accused of committing fraud, tragically illustrate what can happen when public authorities focus on general policies without considering citizens' needs and circumstances.

This chapter argues that promoting empathy as a public value can help achieve a delicate balance between generalization and particularism. This is especially valuable at a time when technology can be used either to standardize, guarantee consistency, and reduce 'noise'³ or to personalize services to individuals' needs.⁴ While personalization of advertisement and services is better known in the world of consumer goods and services, there is also margin to fit personalization in the public sector, namely through empathy.⁵ Empathic approaches can respond to circumstances of acute vulnerability and ensure that we strike the balance between the required generalizations and the particularized approach that can ensure that individuals' needs are considered. This has been increasingly important with the digitalization and automation of government and administrative decision-making.

With the increasing automation and digitalization of governments in recent years there has been a growing call, particularly in the Netherlands, from both scholars

¹ See Goanta (2022).

² See Ranchordás and Scarcella (2021).

³ See Kahneman et al. (2021).

⁴ See Ben-Shahar and Porat (2021).

⁵ See Coglianese (2022).

and policymakers to emphasize the individual dimension of citizens.⁶ In the age of automation, citizens often find themselves overlooked and reduced to categories by algorithms that fail to recognize their unique backgrounds, challenges, virtues, and circumstances. When they have direct contact with human civil servants, they are often confronted with bureaucratic systems, reduced assistance, and limited compassion for their challenges. As a result, citizens feel unseen and misunderstood by governments and their policy tools.

This chapter argues for the inclusion of empathy as a public value in the toolkit of public administration. Empathy adds to existing public values because it has explanatory power: it allows public authorities to understand how to pursue the public interest in a way that fits individual challenges. Being able to empathize means being able to recognize what the other is experiencing and respond to it. While there are many definitions of ‘empathy’, Baron-Cohen defines it as follows:

“Empathy is our ability to identify what someone else is thinking or feeling, and to respond to their thoughts and feelings with an appropriate emotion.”⁷

Empathy is thus connected to compassion which involves recognizing someone’s suffering, responding emotionally to it, experiencing empathic concern, and having the desire to alleviate it.⁸ Empathy differs from altruism, the wish to help others. Empathy, in this chapter, is understood primarily as a behavioral response, rather than a feeling. Empathic behavior shows an understanding of how heavy or tight the other person’s shoes are, thereby grasping why a person is walking in such a way. Could it be that this person’s shoes are so tight that walking straight has become impossible? Is this why this person cannot run, but only limp? Empathy as a public value does not mean that civil servants should feel sorry for citizens who find themselves in vulnerable conditions, even though feelings and behaviors may coincide. Instead, empathy here involves responding with appropriate measures while ensuring good governance, providing accessible and adapted communication, and training civil servants to better assist citizens in challenging circumstances, regardless of their personal beliefs.⁹

Our argument proceeds in four steps. First, we discuss vulnerability as the central problem of citizens’ interactions with digital government. Digital government often assumes that citizens should be able to independently engage with public services.¹⁰ However, this gives rise to (administrative) vulnerabilities for those who are unable to do so. Drawing on socio-legal and legal feminist scholarship, we theorize vulnerability as a phenomenon that is universal and inherent (1), situational (2), multilayered and contextual (3), inflicted by power asymmetries and dependency (4), and imposed externally through the state, society, and its institutions (5). Second, we explain the value of vulnerability in helping us understand the

⁶See Brinkman and Vonk (2022); Heukelom-Verhage (2020).

⁷See Baron-Cohen (2011), p. 12.

⁸See Kanov et al. (2004); Strauss et al. (2016).

⁹See Coglianese (2023).

¹⁰See Meijer (2012).

importance of considering individual features and circumstances. As a third step, we delve into the concept of empathy, which we propose as a public value and a response to administrative vulnerability. Lastly, we consider possible critiques of empathy as a public value.

2 Self-Reliance Before Digital Government and Vulnerability

Since the 1990s, governments have progressively transitioned from a ‘paper-based’ and analogical public administration to a digital government characterized by the increasing digitization of forms and automation of public services.¹¹ Today, ICTs, automated decision-making systems, and algorithms have become integral parts of government.¹² Simultaneously, the internet has become the primary, and sometimes exclusive, infrastructure for applying for welfare benefits, licenses, adjudicating claims, or filing tax returns.¹³ These developments have significantly improved the efficiency of public administration, advanced coherence in administrative decision-making, and substantially reduced government costs.¹⁴ However, the digitalization and automation of government have also produced new vulnerabilities for citizens, often manifested in citizens’ inability to enforce their rights on equal terms or infringements of citizens’ administrative rights, discussed in the following.

2.1 *Self-Reliance in Digital Government*

Research shows that digitalization and automation of the public sector indeed has many positive aspects to it, such as convenience, 24/7 contact with the government, and the simplification of many (but certainly not all) government interactions. Nevertheless, standardized and automated communication does not amount to improved communication with citizens, as many of these policies were primarily conceived to cut costs and operate in the government’s interest rather than to protect citizens’ needs. Citizen-centered design is a much more recent concern for governments that did not drive the original digitalization movements. Instead, digital government was originally designed assuming ‘average skills,’ self-reliance, thus overlooking that different individuals have different abilities. As digital public services are built on ableist structures, they exclude individuals who are incapable of navigating digital government independently. In essence, “ableism reflects the sentiment of certain social groups and social structures that value and promote certain

¹¹ See Ranchordás (2022); Viana (2021); Tomlinson (2019).

¹² See Viana (2021).

¹³ See Ranchordás (2022).

¹⁴ See Ranchordás and Scarcella (2021); Madsen et al. (2022); Schou and Pors (2019).

abilities, such as productivity and competitiveness, over others, such as empathy, compassion, and kindness.”¹⁵ Consequently, ableist systems in digital government hinder the equitable participation of individuals in society who fall outside normative standards.¹⁶

Not surprisingly, digital government has added, for many citizens, a new layer of complexity: digital bureaucracy. While digital forms can vary widely in their complexity and user-friendliness, many of them require considerable time and attention to fill out. Moreover, digital government is often accompanied by complex automated systems that operate like black boxes making it difficult, often impossible, to grasp how the algorithm reaches the decision.¹⁷ Automation has given rise to new vulnerabilities and exacerbated existing ones. These vulnerabilities can worsen existing socio-economic inequalities and have the potential to strain government-citizen relations.¹⁸ In an ongoing project, we qualify this phenomenon as administrative vulnerability.¹⁹

2.2 *Administrative Vulnerability*

The notion of ‘administrative vulnerability’ refers to a set of circumstances in which citizens are unable to enforce their fundamental rights or in which their rights are infringed, often exacerbating asymmetric power relationships between citizens and the government. There are many types of administrative vulnerabilities. While some are specific to digital government and are thus ‘new’, others result from complex bureaucracy or socioeconomic challenges. We divide these vulnerabilities in two dimensions: first, the digitalization dimension, defined by restricted access to digital services (forms, digital platforms); second, the automation dimension, defined by potential rights violations. In the first dimension, citizens may be unable to access welfare services or contest administrative decisions online due to accessibility problems.

In the second dimension, citizens cannot exercise their rights the way in which decision-making is automated and how digital technology is used for enforcement. Therefore, vulnerability in this dimension arises from citizens’ increased susceptibility to rights infringement, which may result from biased and discriminatory data, opaque decision-making practices, and issues related to transparency and accountability. In the next section, we discuss a vulnerability framework that can help us better understand these dimensions of administrative vulnerability.

¹⁵ See Wolbring (2008), p. 253.

¹⁶ See Campbell (2013).

¹⁷ See Busuico (2021).

¹⁸ See Schou and Pors (2019).

¹⁹ Project website: Vulnerability in the Digital Administrative State | Tilburg University (funded by the Dutch Research Council, NWO).

3 Vulnerability Framework

This section justifies the need to draw on vulnerability theories and how these theories can help us understand administrative vulnerability in relation to digital government.

3.1 *Why This Framework?*

The concept of vulnerability offers a meaningful framework at three levels. Firstly, vulnerability enables an appreciation of the human perspective in government, especially at a time when citizens are increasingly reduced to digits and data points. Thus, the concept bears the potential to divert attention back to the individual, acknowledging the unique challenges and exposure to harm within digital government. This is particularly important in an era where citizens risk being dehumanized and reduced to statistical and homogenous entities, as it allows them to recognize their diverse circumstances and traits.

Secondly, vulnerability can serve as a tool to highlight social disparities and social justice issues.²⁰ As vulnerability may manifest through limitations on citizens' rights, it holds significant potential to reveal how digital government obstructs individuals from asserting their rights.²¹ Furthermore, unveiling social justice issues and understanding citizens as vulnerable implies a shared obligation and ethical and collective responsibility to protect those at risk from harm.²² When the state directly, yet presumably unintentionally, inflicts vulnerabilities, its role in mitigating vulnerability becomes even more pronounced.²³

Finally, vulnerability, as a conceptual framework, serves as a relevant instrument for delineating the individuals or groups susceptible to harm, shedding light on the institutional and systemic factors underpinning these vulnerabilities.²⁴ Thus, vulnerability facilitates a proactive stance, enabling the identification of precursory risk elements associated with potential harm arising from engagements with digital governance.

²⁰ See Mackenzie et al. (2014); Heri (2021); Fineman (2018, 2021).

²¹ Hogan and Marandola (2005), p. 459.

²² See Fineman (2021).

²³ Heri (2021), p. 22.

²⁴ See Schroeder and Gefenas (2009).

3.2 Meaningful Theoretical Foundations in the Context of Digital Government

This section delineates five theoretical approaches to vulnerability that provide meaningful frameworks to understand vulnerability in the context of digital government. Thus, vulnerability can be understood as a phenomenon that is universal and inherent (1), situational (2), multilayered and contextual (3), inflicted by power asymmetries and dependency (4), and inflicted externally through the state, society and its institutions (5).

3.2.1 Vulnerability as a Universal and Inherently Human Phenomenon

Firstly, vulnerabilities of citizens are universal due to the inherently vulnerable nature of all legal subjects.²⁵ Fineman's interpretation of vulnerability encompasses a constant exposure to risk of harm, representing a persistent threat that is often beyond human control that cannot be entirely eliminated.²⁶

Her legal feminist universal vulnerability paradigm contrasts the particular approach to vulnerability, which limits vulnerability to specific population groups. This particular understanding of vulnerability used to be the prominent comprehension of vulnerability, also in its legal application. However, it has received increasing criticism in recent years due to its risk to stereotype and stigmatize entire population groups. Furthermore, despite the label's blame related sentiments and common associations with 'weakness' and 'victimhood', the particular vulnerability approach bears the risk of overlooking individuals who fall outside these common vulnerability categories but are yet vulnerable.

Fineman's idea of the inherently dependent and universally vulnerable legal subject not only contents that vulnerability is inherent to the human condition (ontological) but also argues that a strong and responsive state is required to mitigate vulnerabilities of its citizens to build resilience.²⁷ This is particularly interesting in the context of digital government the state itself involuntarily inflicts vulnerability on its citizens, making its role to combat them even more pronounced. Moreover, Fineman criticizes that today's neoliberal state excessively emphasizes the autonomy and independence of the liberal subject.²⁸ The digitalization and automation of digital government mirrors this: individuals are expected to be self-reliant and carry the burden of bureaucracy with little assistance. While Fineman's idea of the universal and dependent subject can be criticized based on its difficulty to be operationalized, risk for paternalistic policy structures and lack of the value of autonomy, it

²⁵ Fineman (2008), p. 1.

²⁶ Fineman (2008), p. 9.

²⁷ See Fineman (2010).

²⁸ See Fineman (2010).

offers an important departure from particularized vulnerability notions to an understanding that everyone is at constant risk of harm.²⁹

3.2.2 Vulnerability as a Situational Phenomenon

Secondly, despite the universal nature of vulnerability in the context of digital government, vulnerability is also a situational phenomenon. Mackenzie (2014) claims that situational forms of vulnerability are more context-specific and “caused or exacerbated by social, political, economic or environmental factors.”³⁰ Thus, understanding vulnerability as a situational phenomenon in digital government sheds light on specific circumstances, conditions, or contexts that contribute to the individual or a group. Notwithstanding, understanding vulnerability in digital government as temporary or contingent upon particular situations allows for a meaningful detachment of the individual’s characteristics to a focus on the circumstances in which institutional (legal) practices give rise to vulnerabilities for citizens. In the context of digital government, situational vulnerabilities can reach from algorithmic discrimination and inadequate redress mechanisms to challenge administrative automated decisions to limited digital literacy and technological barriers, such as a poor internet connection.

Furthermore, it is crucial to acknowledge that inherent and situational vulnerabilities can be interlinked and reinforce each other.³¹

3.2.3 Vulnerability as a Multi-layered and Contextual Phenomenon

Third, understanding vulnerability as a multi-layered and elusive phenomenon provides a meaningful account in the context of digital government. Florencia Luna advocates for the contextuality and complexity of vulnerability by promoting the notion that vulnerability is composed of many different layers.³² Similar to Fineman, this approach disavows the particular approach to vulnerability, also not limiting it to specific population groups. Luna stresses that investigating how each layer contributes to vulnerability is possible without enforcing a rigid hierarchy of these layers.³³ Furthermore, Luna suggests that different social contexts can result in various experiences of vulnerabilities for individuals, depending on the degree, context, accumulation and interplay of vulnerability layers.³⁴ This underlines the dynamic,

²⁹ See Kohn (2014); Mackenzie et al. (2014); Brown et al. (2017).

³⁰ Mackenzie (2014), p. 39.

³¹ Mackenzie et al. (2014), p. 39.

³² See Luna (2009, 2019).

³³ See Luna (2009).

³⁴ See Luna (2009).

situational and fluid nature of vulnerability, making it a condition rather than a fixed status.³⁵

This intersectional account of vulnerability also sheds light on vulnerable legal subjects that are not commonly associated with belonging to a vulnerable population group. For instance, a man from the upper middle class is not commonly considered vulnerable. However, layers that are not visible at first sight, such as loneliness, mental health struggles or having gone through a bureaucratically burdensome divorce, could render him vulnerable when interacting with digital government. Conversely, a citizen with a documented learning disability might have other compensating layers such as a good socio-economic background or a supporting partner or family member, thus mitigating the possible vulnerability inflicted by the impairment.

In sum, Luna's layered approach to vulnerability provides a meaningful theoretical framework in the context of digital government as it favors the application of a universal and non-exclusive lens on vulnerability while allowing for an operationalization of the concept and acknowledging the elusive and accumulative effect of vulnerability layers within different contexts.

3.2.4 Vulnerability as a Phenomenon Inflicted by Power Asymmetries and Dependency

Fourth, contemporary citizen-government relations in public administration inherently feature power asymmetries, wherein governments possess the unilateral authority to define individuals' positions and exercise public authority.³⁶ These asymmetric power dynamics elevate the risk of harm for citizens, ultimately giving rise to vulnerability.³⁷ While new communication channels have become available to citizens, facilitating their interactions with public authorities, the increasing reliance on self-service mechanisms reshapes the traditional power dynamics between citizens and the government. Consequently, the digitalization of governments paradoxically both connects and disconnects citizens from government.³⁸ Increasing power imbalances, particularly in the context of citizens' reduced ability to legally contest automated decisions made against them,³⁹ becomes increasingly evident. This inability to access legal remedies and challenge injustices significantly exposes citizens to vulnerability.

³⁵ See Luna (2009).

³⁶ See Ranchordás (2022).

³⁷ See Zarowsky et al. (2013).

³⁸ Lindgren et al. (2019), p. 430.

³⁹ See Tomlinson (2019).

3.2.5 Vulnerability as an Externally Inflicted Phenomenon

Finally, a prevalent pitfall of vulnerability lies in attributing blame or responsibility onto the citizen that is risk of instead of the state, society and its institutions.⁴⁰ Fineman advocates for a reorientation of responsibility for vulnerability and claims that the individual is never responsible for contesting their differences, as a disadvantage is always produced externally.⁴¹

Pathogenic vulnerability, categorized as a form of situational vulnerability, embodies an external perspective by highlighting detrimental factors originating from the environment or systemic issues that render citizens vulnerable.⁴² Such pathogenic vulnerabilities arise from “prejudice or abuse in interpersonal relationships and from social domination, oppression, or political violence.”⁴³ To employ an example in the context of digital government, biased discriminatory data fed to an algorithm is grounded in discriminatory societal and institutional sentiments. Individuals who are flagged by a predictive algorithm based on such bias are thus subject to pathogenic vulnerabilities. Therefore, vulnerability does arise due to the individual’s attributes that might deviate from societal norms but to existing societal sentiments. External, vulnerability inflicting parties are then responsible for combatting and mitigating the vulnerabilities they produce.⁴⁴ However, it is vital to be careful not to concede the assumed needs of vulnerable population groups, as this can give rise to new pathogenic vulnerabilities and paternalistic policy structures.⁴⁵

These vulnerability frameworks help understand citizen-government dynamics and acknowledge challenges that citizens encounter in their engagement with the public administration. Empathy conceived as a public value, discussed in the next section, can offer a possible response to this problem.

4 Empathy as a Public Value

In civil law traditions of administrative law, public authorities are presented as institutions that exist for the sole purpose of pursuing the public interest. In doing so, these institutions should consider public values, that is, certain fundamental ideals, principles or even ethical considerations that are at the core of a society and guide government. Examples are autonomy, fairness, transparency, privacy, and

⁴⁰ See Brown et al. (2017).

⁴¹ Fineman (2012), p. 638.

⁴² Mackenzie (2014), p. 39.

⁴³ Mackenzie (2014), p. 39.

⁴⁴ Mackenzie (2014), p. 40.

⁴⁵ Mackenzie (2014), p. 40.

sustainability. Public values are ‘criteria for action’⁴⁶ which inform not only the attitudes of public servants but also the ends to which their public authorities should aspire.

However, as Bozeman explains, values are not based on the needs of society but on the individual needs of humans as such. Government, given its position of power and, in the case of local governments, its proximity to citizenry, is well-placed to address the needs of individuals and thus to further them through the pursuit of public values.⁴⁷ These values serve as guiding principles for the governance of public authorities. Jørgensen and Bozeman have argued that while public service values exist within a hierarchy and are causally related, we cannot elect one single value as being the leading and central public value of a government.⁴⁸ Other scholars in the field support this position.⁴⁹

While the notion of public values varies depending on the society or community, key public values often include, for example, accountability, which, like empathy, can be seen as a moral virtue or a mechanism to render public officials answerable for their actions. Other common public values encompass autonomy, privacy, transparency, fairness, and efficiency. Public values are often closely related to human and fundamental rights since these also embody public and social values. However, when arguing that digital government should be designed for certain public values, we go beyond established legal frameworks as we seek to understand what values lie underneath. This allows us to make the claim in this section that also empathy should be considered as a public value in the context of public administration, as it is a valuable tool to promote good administration, balance power asymmetries, and innovate government.

4.1 Conceptualizing Empathy as a Public Value

Empathy has been studied by a large number of scientific fields (e.g., psychology, management studies, ethics) and is a term that encompasses multiple dimensions and applications. From an ethical perspective, empathy is often connected to other virtues that are part of ‘being a good citizen,’ ‘being sensitive,’ and a ‘good person’. However, the idea of having empathy as a public value (rather than as mere virtue) is less straightforward, if not, controversial.

Empathy is commonly used to refer to different realities and behaviours. Phrases such as “you are not empathetic” can be perceived as insulting. Empathy is often equated with being nice, kind, charitable, and understanding. Women are more likely to be described as ‘empathetic’, a stereotype which might not necessarily be

⁴⁶ See Molina (2009); Denhardt and Denhardt (2015).

⁴⁷ See Bozeman (2007).

⁴⁸ See Jørgensen and Bozeman (2007).

⁴⁹ See Box (2015); Molina and McKeown (2012).

backed by scientific support, depending on how you define empathy. This gendered view stems from the notion that ‘empathy and emotional sensitivity’ are female traits, which also perpetuates the assumption that women are less rational than men.⁵⁰ There are thus multiple definitions and types of empathy, many of which do not entail any inner feeling of sorrow for others or kindness.

Edlins and Dolamore define empathy in the public realm as “the ability to recognize, understand, and respond to the feelings of another [offering] a way to improve [public servant-citizen] interactions and bring them more in line with expected public service values.”⁵¹ Empathy includes four key elements: (1) the ability to see the world as others’ see it, (2) to understand another’s current feelings, (3) to remain non-judgmental, and (4) be able to communicate understanding of another’s feelings.⁵² While these elements are important to understand empathy in broad terms, Dolamore differentiates between ‘stepping into another’s shoes’ and reacting to this emotional state.⁵³ For empathy to occur, the information obtained from the imaginative process must be utilized to communicate understanding back to the other individual and to generate a reaction in line with the information obtained.⁵⁴ In this way, empathy starts as an internal process and culminates with an active response. The combined layers of an internal process and an active response is also what makes empathy different from sympathy, and other one-dimensional activities which are comprised in ‘pseudo empathy’ or a ‘self-oriented perspective taking’.⁵⁵

Empathy can be thus rational (*recognizing the need to reason differently*), emotional (*feeling empathy*) or behavioral (*acting in an empathetic way*). As an emotional skill, empathy means that a person can truly understand or feel the situation of another.⁵⁶ Emotional empathy can also encompass “the ability to recognize, perceive, feel, and act on the emotional state of another person.”⁵⁷ While the emotional and behavioral dimensions may overlap, they also differ greatly. The passive ability to observe and appreciate the other differs considerably from the requirement to feel or act upon someone else’s emotional state. References to empathy in the public administration literature suggest it is a fuzzy and context-dependent concept. For many recipients of empathy, empathetic behavior may suffice to make them feel ‘seen and heard,’ even if the administrative decision remains unchanged. This is because empathy primarily focuses on the experience rather than the outcome. Thus, incorporating empathy as a public value is not necessarily about altering the outcomes of citizen-government interactions but about enhancing the experience to better meet citizens’ needs and promote values such as fairness and equality.

⁵⁰ Gilligan (2014), p. 96.

⁵¹ Edlins and Dolamore (2018), p. 301.

⁵² Edlins and Dolamore (2018), p. 302.

⁵³ See Dolamore (2021).

⁵⁴ See Wiseman (1996).

⁵⁵ Coplan (2011), p. 40.

⁵⁶ See Guy et al. (2008).

⁵⁷ See Batson (2009) cited in: Zanetti and King (2013).

In public service, the concept of empathy has emerged as a critical pillar for effectively navigating the complexities of discretion while upholding core public values.⁵⁸ Empathy transcends the boundaries of bureaucracy and resonates deeply with the principles of fairness, equity, and the responsibility of government agencies to serve all citizens. Empathy is also listed as one of the many public service values.⁵⁹ Guy, Newman, & Mastracci (2008) define empathy, in the context of public service, as “the ability to understand and appreciate the citizen’s situation and respond accordingly,” but also suggest that it is usually paired with the active suppression of street-level bureaucrat’s “own emotions while simultaneously expressing an alternate emotion.”⁶⁰

As a public value, empathy can serve as the bridge between the institutional machinery of government and the lived experiences of the people it serves. When public servants practice empathy in their interactions and decisions they can “make compensatory adjustments for known distinctions between self and other.”⁶¹ In this context, empathy can simply consist of small changes to public action such as using empathetic communication that expresses appreciation for the citizen’s circumstances or voices sympathy for the negative outcome of the result. For example, for several years, Dutch researchers applying for grants offered by the Dutch Science Organization (NWO) would receive an email stating that “if their application number was not among the list of the awarded grants, they had not been selected for funding.” Applying for scientific funding is a stressful and time-consuming endeavor that can be career changing.⁶² Although it is obvious that this public authority cannot award grants to all researchers and its human resources are limited, NWO could have adapted its communication to those individuals that had made serious efforts but were not funded. In the meantime, NWO has indeed done so. Communication with scientists has improved during and immediately after the pandemic, and researchers were, for instance, able to disclose if this extraordinary period had affected their careers. This measure, even though unlikely to change the course of the grant evaluation process, can be seen as a step towards more empathetic communication with citizens. Sentences such as “we appreciate the effort” would probably be appreciated by the researchers whose proposals are rejected.

The call for greater empathy within the public service has echoed not only in academic circles but also in the highest echelons of government. In the United States, President Barack Obama and judge Sonia Sotomayor have, for example, underlined the critical role of empathy in the case law of the Supreme Court.⁶³ Their discussions shed light on the notion that empathy is not merely a personal quality

⁵⁸ See Molina and McKeown (2012).

⁵⁹ See Box (2015); Lewis and Gilman (2012).

⁶⁰ Guy et al. (2008), p. 66.

⁶¹ Steinberg (2014), p.50

⁶² While the researchers authoring this paper and the book in which it will be published, are funded by the Dutch Science Organization, this example is entirely independent from the PI’s experience with NWO grant applications.

⁶³ See Baker (2010).

but a fundamental requirement for those entrusted with the responsibility of interpreting and upholding the law.

Furthermore, empathy is a public value that can be used to ‘tame politics,’ as exemplified by the 2013 government shutdown in the United States, where elected officials in Congress display a strong lack of empathy toward their constituents.⁶⁴ Empathy, as a public value, can contribute to good administration, avoiding that the suffering of the public is forgotten due to political or technocratic pressures. Furthermore, empathy can contribute to a solution to these vulnerabilities created by the digitization of government, bridging the digital divide and ensuring that the human element remains at the core of governance.

4.2 *Empathy as Good Governance*

Empathy can be understood as a public value to further good governance and, in so doing, ensure that public authorities also promote the principles of good administration. We focus on a selected number of aspects of good governance: innovation, professionalism, effective communication and meaningful interaction with government.

4.2.1 Innovative Governance

Empathy can be used to promote innovative governance and reshape design thinking within government. Design thinking is a framework that places empathy at its core within the innovation process.⁶⁵ This approach, defined as “an array of mind-sets, methods, and practices to help people become more productive, creative, and innovative,” is considered a powerful tool for addressing complex problems and meeting customer needs.⁶⁶

Moreover, Chowdhury and Beresford have delved into the transformative power of empathy within the context of innovation in government in Bangladesh.⁶⁷ Their work showcases how the Access to Information (a2i) program utilizes empathy to fuel progress and creativity. Their exploration highlights the strategic deployment of empathy as a catalyst for fostering innovation and change, offering valuable insights into how empathetic approaches can drive positive societal transformations in diverse settings. In the context of Bangladesh’s public administration, a remarkable transformation has been underway through the Empathy Training Programme (ETP), an initiative designed to instill empathy among street-level bureaucrats. This

⁶⁴ See Congressional Record (2013).

⁶⁵ See Clarke and Craft (2019); McGann et al. (2018); Lewis et al. (2020).

⁶⁶ Liedtka (2013); see Roper et al. (2016); Pfothenhauer et al. (2019).

⁶⁷ See Chowdhury and Beresford (2017).

program has successfully reshaped the once-prevalent paternalistic and colonial mindset that characterized public service. At the heart of the ETP methodology lies a powerful approach: senior public servants assume the role of ‘mystery shoppers’, stepping outside their ministry or area of expertise to engage with citizens’ access points for services. This immersive experience places them squarely in ‘citizens’ shoes,’ compelling them to navigate the intricate web of public systems without any official or intellectual privileges. Through this first-hand encounter, participants cultivate a discerning and empathetic perspective that they subsequently apply to scrutinize their own agency’s delivery systems.

4.2.2 Empathy as Caring, Meaningful, and Effective Communication

Citizen-friendly communication as well as the duties of care and diligence are essential elements of good administration or good governance. Empathic communication within public organizations is particularly important to build and sustain trust in government, particularly at times of crisis. This has also been explained from an organizational theory perspective. According to this theory, crisis management performance is intrinsically linked to governance capacity and governance legitimacy.⁶⁸ The current backdrop of crises, particularly those witnessed in 2020–2022, has accompanied a period of unprecedented uncertainty and transboundary challenges, significantly impacting the governance capacities of many public organizations. Thus, improved citizen-government relations through the implementation of empathy as a public value would also be meaningful to improve communication between government and citizens but also within public organizations.

During crises such as the COVID-19 pandemic and the ongoing uncertain moments resulting from the wars in Ukraine and Israel, the sense of care and responsiveness that individuals experience are equally important as the services themselves. This intrinsic human need underlines the significance of maintaining meaningful, people-to-government relationships, particularly during crises, as advocated by Christensen et al.⁶⁹ In the words of Guy, “for governing to be effective, citizens must feel good about those who govern and how governing happens.”⁷⁰ Whether grappling with a public health crisis or navigating the complexities of a social-political crisis, people seek not only to witness a response but to feel that their concerns are being acknowledged. In this light, empathy emerges as a symbol in the quest for both effective governance and the preservation of the essential bonds that tie public organizations and the citizens they serve. The citizen-state interaction is central in public administration. Edlins and Dolamore have explained how showing empathy can facilitate meaningful engagements between public officials and

⁶⁸ See Christensen et al. (2016).

⁶⁹ See Christensen et al. (2016).

⁷⁰ Guy (2019), p. 4.

citizens.⁷¹ Empathy enables these interactions to genuinely understand citizens' needs and aspirations. This perspective particularly resonates with adept administrators who have drawn from the organizational development literature, incorporating concepts like teambuilding, motivational programs, and leadership development. Thus, organizational development has long recognized the fundamental necessity for individuals to sense that they are genuinely understood in their interactions.⁷²

Furthermore, Ponomariov and McCabe challenge the conventional dichotomy between professionalism and empathy in public service, arguing that empathic communication should be regarded as professional.⁷³

4.2.3 Equal Treatment, Fairness, and Justice

Empathy can also be interrelated to other public values and fundamental rights such as fairness and equal treatment. Contemporary discourse among scholars widely supports the view that treating all citizens equally can perpetuate rather than alleviate inequality.⁷⁴ Emphasizing on substantive equality, Fineman advocates for a departure from a formal (or legal) equality model, which mainly focuses on an anti-discrimination mandate, to an approach in which substantive equality is achieved by addressing underlying unjust power and opportunity structures that contribute to inequalities.⁷⁵ This means that sometimes unequal treatment is required to achieve equitable outcome and equal opportunity, as some individuals are subject to structural disadvantage and impaired opportunity structures.⁷⁶ Thus, the focus lies in the equality of outcomes and opportunity instead of the equality of treatment.⁷⁷

Aligned with these perspectives are conceptualizations of justice, such as Aristotle's theory distinguishing between compensatory and distributive justice.⁷⁸ While compensatory justice refers to the absolute equality between goods, distributive justice refers to the need for an equal outcome, which might require unequal treatment. Furthermore, administrative justice refers to just outcomes in administrative decision making and is comprised of procedural fairness and substantive justice.⁷⁹ Thus, the effect of decision making, which determines whether a benefit or harm has occurred to the individual is vital in determining administrative justice.⁸⁰

⁷¹ See Edlins and Dolamore (2018).

⁷² Edlins and Dolamore (2018), p. 5.

⁷³ See Ponomariov and McCabe (2017).

⁷⁴ See Braveman et al. (2018); Berry (2016); Shakespeare (2006); Fineman (2014).

⁷⁵ Fineman (2012), p. 638.

⁷⁶ Fineman (2008), p. 16; Fineman (2014), p. 626.

⁷⁷ Fineman (2014), p. 619.

⁷⁸ See Radbruch (1932).

⁷⁹ Adler (2003), p. 324.

⁸⁰ Adler (2003), p. 324.

Consequently, when the public administration lacks empathy and fails to comprehend underlying disadvantages, administrative justice is compromised.⁸¹

Lastly, empathy can also be required by the duty of care or diligence, an important principle of good administration, codified both at national and European level.⁸² Empathy can help ensure that public authorities fulfill the duty of care in a better way, providing explanations that are adapted to the citizen standing in front of the public authority and their specific circumstances.

The automation of governments is often justified by the perceived advantages of algorithmic consistency, alleged objectivity, and rationality in automated systems. However, their inability to account for individual circumstances limits the empathetic assessments of citizens' unique positions. The seemingly fair allocation of resources falls short of achieving substantive equality because some individuals require human judgment, contextual evaluations, and consideration of their specific circumstances due to their vulnerable positions resulting from societal power and opportunity structures. In this context, empathy emerges as a crucial tool to secure equitable outcomes beyond the mere guarantee of equal treatment.

5 Empathy: Potential Shortcomings

We have pointed out the potential of empathy as a public value, especially at a time when the digital government's human dimension is diminishing. However, as these notions can be subject to critique, this section draws attention to potential shortcomings of making empathy a key public value of digital government.

Edlins contends that within the domain of public administration, a notable disconnection exists concerning empathy.⁸³ Although it is often portrayed as a desirable quality, there is a lack of agreement regarding its precise definition and how to practically apply it within administrative settings. Moreover, despite empathy's inherent importance to the field, it has neither been explicitly recognized nor explored as a dedicated research focus in its own right. We offer here an overview of commonly outlined shortcomings of adopting empathy as a public value.

5.1 *Inconsistencies and Discrepancies*

First, in the context of government transactions, empathy may be seen as a source of potential inequalities, inconsistencies, and discrepancies. When empathy is equated to forgiveness (e.g., forgiving first-time mistakes to citizens who do not

⁸¹ Adler (2003), p. 344.

⁸² Hofmann (2020), p. 87.

⁸³ Edlins (2021), p. 3.

master the national language), it could open the door to new inequalities and abuses. While *ignorantia juris non excusat*⁸⁴ is perhaps an unempathetic principle, it guarantees legal certainty and formal equal treatment. At the same time, in low-trust societies, where citizens associate government with abuses of power and excessive discretionary powers, empathy may easily be associated with yet another instrument to disguise irregularities, nepotism and maladministration. Furthermore, empathetic decision-making can be quickly equated with emotional or opinion-based decision-making which is not evidence-based analysis and thus increases the risk of mistakes, inaccuracies, and it could undermine administrative legal systems in the long run. Inconsistencies could also arise when seeking to determine who is more deserving of empathetic treatment. This would entail an ethical assessment, which may be a difficult task for civil servants to take on, ultimately possibly resulting in new injustices. However, despite the risk of inconsistencies and discrepancies, our proposal for empathy as a public value rather depicts a broader idea of empathy that guides public decision making and policy choices instead of being applied in a narrow and single-case manner.

Furthermore, the emphasis on the individual can challenge the feasibility of digital government and its efficiency. Due to significant government cuts over recent decades, many public service areas are understaffed, leaving public servants without the capacity for empathy in their decision-making. Overwhelmed by heavy workloads and limited resources, public servants may seek to expedite decisions, resorting to simplifications and standard policies. While these methods streamline workflow, they can foster bias and ‘othering,’ ultimately undermining empathy.

5.2 *Prioritizing Individual Cases, Missing the Systemic Ones*

Second, by prioritizing individual cases, the risk of failing to address and reveal more significant systemic issues arises. This may boil down to the fact that the broader perspective is overlooked. Systems approaches in public administration have gained increasing popularity as they allow “governments to confront problems that traverse administrative and territorial boundaries in a holistic manner.”⁸⁵ Thus, the application of a systemic lens to complex and wicked problems helps successfully transform public systems.⁸⁶ However, focusing on individual levels can make it challenging to discover underlying systemic issues, and tackling those without a holistic, generalized approach can be problematic.

⁸⁴ Latin for ‘ignorance of the law is no excuse’, meaning that unawareness of the law does not justify its violation.

⁸⁵ OECD (2017), p. 9.

⁸⁶ OECD (2017), p. 11.

5.3 *Empathy as Patronizing*

Third, the excessive use of empathy in digital government can have patronizing effects on citizens. While we have clarified the benefits of empathy in this paper, its establishment as a public value bears the risk of disempowering individuals who are vulnerable before digital government. Interestingly, Samani and colleagues⁸⁷ found a positive correlation between paternalism and empathetic concerns. Thus, they empathize that acting upon empathy can have effects that are opposite to what is intended, undermining individual autonomy.⁸⁸ Similarly, Reeves claims that when focusing empathy on specific population groups, such as racial minorities, there is a high risk of paternalism.⁸⁹ However, while empathy might limit individual autonomy in some circumstances,⁹⁰ it has great potential to overcome vulnerabilities in today's digital government by helping navigate differences between citizens.

5.4 *Empathy Fatigue*

Finally, empathy as a public value can result in the emergence of 'empathy fatigue', which relates to challenges encountered by frontline workers and civil servants. Stebnicki defines the concept as follows: "empathy fatigue results from a state of emotional, mental, physical, and occupational exhaustion that occurs as the counselors' own wounds are continually revisited by the client's life stories of chronic illness, disability, trauma, grief, and loss."⁹¹ Thus, it can be emotionally draining for decision-makers and public servants to apply an empathetic lens to citizens' situations and circumstances. Over time, this exhaustion might impair their ability to sustain compassionate decision-making and impact their mental health.

6 Conclusion

How do we know how it feels to walk in someone else's shoes if we have never worn shoes that are too tight or too loose? By trying on shoes we have never considered, we embrace the diversity of experiences within our society and seek to understand who citizens truly are, beyond mere data points. Empathy ensures that public bodies recognize when they are placing citizens in vulnerable circumstances and hopefully helps them remedy the situation. When vulnerability results from a set of

⁸⁷ See Samani et al. (2022).

⁸⁸ Ibid.

⁸⁹ See Reeves (2017).

⁹⁰ Slote (2004), p. 308.

⁹¹ Stebnicki (2007), p. 317.

complex circumstances, empathy, as a public value, enables government to adapt its discourse and actions to the humans it serves, leading to better acceptance of administrative decisions and policies.

Vulnerability theories promote a meaningful appreciation of the human perspective in today's digital public administration and contribute to our understanding of citizens' unique challenges with government. Referring to the rich body of literature on vulnerability, we built a theoretical framework for understanding vulnerability in the context of digital government based on five approaches. First, we concede with Fineman's universal approach to vulnerability and the notion of the inherently vulnerable and dependent legal subject.⁹² This allows us not to overlook the needs and struggles of citizens who do not fall into common vulnerability categories and simultaneously enables moving away from particularizing, victimizing and stereotyping vulnerability approaches. Second, we base our theoretical understanding of vulnerability on the notion that it can be situational, making it context-specific and dependent on various other factors. Third, we apply a layered and intersectional lens to vulnerability, underlining the concept's multi-dimensional and elusive nature. Thus, the interplay of non-hierarchical layers can exacerbate or mitigate the vulnerabilities of citizens in the context of digital government. Fourth, we concede with the notion that asymmetric power relations and dependencies inflict vulnerability. Thus, asymmetric citizen-government relations trigger administrative vulnerability. Notwithstanding, citizens who are highly dependent on government services have a higher risk of being harmed by digital public administration. Finally, we claim that administrative vulnerability is an externally inflicted phenomenon, theoretically founded on pathogenic vulnerability. In this context, unequal opportunity and power structures are critical vulnerability drivers.

While administrative vulnerability depicts a new problem of today's digitalized and automated public administration, establishing empathy as a public value provides a meaningful pathway to address this issue. This chapter argued that empathy within the public administration bears the potential to significantly enhance governance practices. Its integration can foster innovation and prompt a reevaluation of existing systems that perpetuate inequality. By embracing empathy as a core public value, administrations can create more compassionate, meaningful, and impactful communication channels, thereby improving citizen-government relationships. These positive interactions have the potential to restore trust in public institutions. Moreover, it would bring back a human dimension to a public administration that is characterized by increasing datafication and automation. In this context, this chapter contends that empathy can serve as a powerful instrument in advancing substantive equality and establishing equitable opportunity frameworks. Tailored approaches to address the unique circumstances of individuals can ensure fairness and inclusivity within governance structures. In a landscape dominated by efficiency and consistency, prioritizing empathy offers a pathway to rehumanize public administration and highlights its commitment to serving the diverse needs of society.

⁹² See Fineman (2008).

Despite potential shortcomings, empathy remains a valuable concept as it offers various meaningful opportunities to digital government to improve interactions with citizens, reduce vulnerability sources or compensate for existing inequalities and burdens. Moreover, many of those points of criticism come down to a micro-level application of empathy. In this chapter, we have argued that empathy should be a guiding value in today's digital public administration as a whole, informing policy choices and allowing for tailored assessment when required. Empathy allows a more humanized and positive, compassionate and considerate attitude from the government to the citizen, ultimately improving government-citizen dynamics and combatting administrative vulnerability.

Acknowledgements We thank the editors of this volume, especially Antonia Stanojević for their insightful comments and Madalina Nicolai for research assistance.

References

- Adler M (2003) A socio-legal approach to administrative justice. *Law Policy* 25(4):2–352
- Baker P (2010) In Court Nominees, Is Obama Looking for Empathy by Another Name?, *New York Times*, <https://www.nytimes.com/2010/04/26/us/politics/26memo.html>
- Baron-Cohen S (2011) *Zero degrees of empathy. A new theory of human cruelty and kindness*. Penguin Books
- Batson CD (2009) These things called empathy: Eight related but distinct phenomena.
- Ben-Shahar O, Porat A (2021) *Personalized law: different rules for different people*. Oxford University Press.
- Berry J (2016) Diversity and equity. *Cross Cult Strateg Manag* 23(3):413–418
- Box R C (2015) *Public service values*. Routledge
- Bozeman B (2007) *Public values and public interest: counterbalancing economic individualism*. Georgetown University Press
- Braveman P, Arkin E, Orleans T, Proctor D, Acker J, Plough A (2018) What is health equity? *Behav Sci Policy* 4(1):1–14
- Brinkman O, Vonk G (2022) De menselijke maat. *S&D* 79:28–40
- Brown K, Ecclestone K, Emmel N (2017) The many faces of vulnerability. *Soc Policy Soc* 16(3):497–510
- Busuioc M (2021) Accountable artificial intelligence: Holding algorithms to account. *Public Adm Rev* 81(5):825–836
- Campbell FK (June 2013) Ableism: a theory of everything? International Conference on 'Linking Concepts of Critical Ableism and Racism Studies with Research on Conflicts of Participation' University of Hamburg (Germany). https://disabilitystudies.nl/sites/default/files/ableism_a_theory_of_everything_keynote_f.pdf
- Christensen T, Lægreid P, Rykkja LH (2016) Organizing for crisis management: Building governance capacity and legitimacy. *Public Adm Rev* 76(6):887–897
- Chowdhury A, Beresford N (6 February 2017) How a2i is using empathy to foster innovation in Bangladesh. LSE South Asia Centre. <https://blogs.lse.ac.uk/southasia/2017/02/06/how-a2i-is-using-empathy-to-foster-innovation-in-bangladesh/>
- Clarke A, Craft J (2019) The twin faces of public sector design. *Governance* 32(1):5–21
- Coglianesse C (3 April 2022) Moving toward personalized law. University of Chicago Law Review Online. https://scholarship.law.upenn.edu/faculty_scholarship/2819/

- Coglianesi C (2023) Law and empathy in the automated state. In: Bednarz Z, Zalnieriute M (eds) Money, power, and AI: automated banks and automated states. Cambridge University Press, pp 173–188
- Congressional Record (2013). Proceedings and debated of the 113th Congress, first session: House of Representatives. <https://www.gpo.gov/fdsys/pkg/CREC-2013-10-08/pdf/CREC-2013-10-08.pdf>
- Coplan A (2011) Will the real empathy please stand up? A case for a narrow conceptualization. *South J Philos* 49:40–65
- Denhardt J, Denhardt R (2015) The new public service revisited. *Public Adm Rev* 75(5):664–672
- Dolamore S, Lovell D, Collins H, Kline A (2021) The role of empathy in organizational communication during times of crisis. *Adm Theory Praxis* 43(3):366–375
- Edlins M (2021) Developing a model of empathy for public administration. *Adm Theory Praxis* 43(1):22–41
- Edlins M, Dolamore S (2018) Ready to serve the public? The role of empathy in public service education programs. *J Public Aff Educ* 24(3):300–320
- Fineman MA (2008) The vulnerable subject: anchoring equality in the human condition. *Yale J Law Fem* 20(1):1–20
- Fineman MA (2010) The vulnerable subject and the responsive state. *Emory Law J* 60:251–275
- Fineman MA (2012) Feminism, masculinities, and multiple identities. *Nevada Law J* 13(2):619–639
- Fineman MA (2014) Equality and difference: the restrained state. *Alabama Law Rev* 66(3):609–625
- Fineman MA (2018) Vulnerability and social justice. *Valparaiso Univ Law Rev* 53(2):341–369
- Fineman M (2021) Universality, vulnerability, and collective responsibility. *Les ateliers de l'éthique* 16(1):103–116
- Gilligan C (2014) Moral injury and the ethic of care: reframing the conversation about differences. *J Soc Philos* 45(1):89–106
- Goanta C (2022). The Ancient Alien: Good Faith as the Facilitator of Personalized Law. *U. Chi. L. Rev. Online*, 1.
- Guy ME, Newman MA, Mastracci SH (2008) Emotional labor: Putting the service in public service. Armonk, NY: M.E. Sharpe.
- Guy ME (2019) Expanding the toolbox: why the citizen-state encounter demands it. *Public Perform Manag Rev* 44(3):1–18
- Heri C (2021) Responsive human rights: vulnerability, ill-treatment and the ECtHR. Bloomsbury Publishing
- Heukelom-Verhage S, van der Berge L, Vermaat MF, Lurks M, van Renssen NA (2020) Maatwerk in het bestuursrecht. Boom Juridisch
- Hofmann HC (2020) The duty of care in EU public law—A principle between discretion and proportionality. *Rev Eur Adm Law* 13(2):87–112
- Hogan DJ, Marandola E Jr (2005) Towards an interdisciplinary conceptualisation of vulnerability. *Popul Space Place* 11(6):455–471
- Jørgensen TB, Bozeman B (2007) Public values: an inventory. *Adm Soc* 39(3):354–381
- Kahneman D (2021) Noise: a flaw in human judgment. HarperCollins.
- Kanov JM, Maitlis S, Worline MC, Dutton JE, Frost PJ, Lilius JM (2004) Compassion in organizational life. *Am Behav Sci* 47(6):808–827
- Kohn N A (2014) Vulnerability theory and the role of government. *Yale JL & Feminism*, 26:1
- Lewis CW, Gilman SC (2012) The ethics challenge in public service: A problem-solving guide (3rd ed.). New York, NY: Jossey-Bass.
- Lewis P, Stenetorp P, Riedel S (2020) Question and answer test-train overlap in open-domain question answering datasets.
- Liedtka J, King A, Bennett K (2013) Solving problems with design thinking: Ten stories of what works. Columbia University Press.
- Lindgren I, Madsen CØ, Hofmann S, Melin U (2019) Close encounters of the digital kind: a research agenda for the digitalization of public services. *Gov Inf Q* 36(3):427–436

- Luna F (2009) Elucidating the concept of vulnerability: layers not labels. *Int J Fem Approaches Bioethics (IJFAB)* 2(1):121–139
- Luna F (2019) Identifying and evaluating layers of vulnerability: a way forward. *Dev World Bioethics* 19(2):86–95
- Mackenzie C, Rogers W, Dodds S (2014) Introduction: what is vulnerability and why does it matter for moral theory. In: Mackenzie C, Rogers W, Dodds S (eds) *Vulnerability: new essays in ethics and feminist philosophy*. Oxford University Press, pp 1–29
- Madsen CØ, Lindgren I, Melin U (2022) The accidental caseworker: how digital self-service influences citizens' administrative burden. *Gov Inf Q* 39(1):1–11
- McGann M, Blomkamp E, Lewis JM (2018) The rise of public sector innovation labs: experiments in design thinking for policy. *Policy Sci* 51(3):249–267
- Meijer A (2012) The do it yourself state. *Inf Polity* 17(3):303–314
- Molina A D (2009) Values in public administration: The role of organizational culture. *Int J Organ Theory Behav* 12(2):266–279
- Molina AD, McKeown CL (2012) The heart of the profession: understanding public service values. *J Public Aff Educ* 18(2):375–396
- Newman M A, Guy M E & Mastracci, S H (2009) *Beyond cognition: Affective leadership and emotional labor*. In *Public sector leadership*. Edward Elgar Publishing.
- OECD (2017) *Systems approaches to public sector challenges: working with change*. OECD Publishing
- Pfotenhauer SM, Juhl J, Aarden E (2019) Challenging the “deficit model” of innovation: Framing policy issues under the innovation imperative. *Res Policy* 48(4):895–904
- Ponomarev B, McCabe B (2017) Professionalism vs. public service motivation: can public administration education alleviate the tension? *Adm Theory Praxis* 39(2):80–99
- Radbruch G (1932) *Gustav Radbruchs Rechtsphilosophie*. Verlag von Queer & Meyer
- Ranchordás S (2022) Connected but still excluded? Digital exclusion beyond internet access. In: Ienca M, Pollicino O, Liguori L, Stefanini E, Andorno R (eds) *The Cambridge handbook of information technology, life sciences and human rights*. Cambridge University Press, pp 244–258
- Ranchordás S, Scarcella L (2021) Automated government for vulnerable citizens: intermediating rights. *William Mary Bill Rights J* 30(2):373–418
- Reeves SL (2017) When compassion leads to paternalism: how empathy can create perceptions of incompetence. *The Ohio State University*
- Roper S, Micheli, P, Love JH, Vahter P (2016) The roles and effectiveness of design in new product development: A study of Irish manufacturers. *Res Policy* 45(1):319–329
- Samani H, Golbabaie S, Borhani K (2022) Predicting paternalism based on components of empathy and behavioral contagion. *Soc Psychol Res* 12(46):101–122
- Schou J, Pors AS (2019) Digital by default? A qualitative study of exclusion in digitalised welfare. *Soc Policy Adm* 53(3):464–477
- Schroeder D, Gefenas E (2009) Vulnerability: too vague and too broad? *Camb Q Healthcare Ethics* 18(2):113–121
- Shakespeare T (2006) The social model of disability. *Disabil Stud Reader* 2:197–204
- Slote M (2004) Autonomy and empathy. *Soc Philos Policy* 21(1):293–309
- Stebnicki MA (2007) Empathy fatigue: healing the mind, body, and spirit of professional counselors. *Am J Psychiatr Rehabil* 10(4):317–338
- Steinberg J (2014) An epistemic case for empathy. *Pac Philos Q* 95(1):47–71
- Strauss C, Taylor BL, Gu J, Kuyken W, Baer R, Jones F, Cavanagh K (2016) What is compassion and how can we measure it? A review of definitions and measures. *Clin Psychol Rev* 47:15–27
- Tomlinson J (2019) How digital administrative justice is made. In: Tomlinson J (ed) *Justice in the digital state: assessing the next revolution in administrative justice*. Policy Press, pp 63–88
- Viana ACA (2021) Digital transformation in public administration: from E-Government to digital government. *Int J Digit Law* 1:29–44

- Wiseman T (1996) A concept analysis of empathy. *J Adv Nurs* 23(6):1162–1167
- Wolbring G (2008) The politics of ableism. *Development* 51(2):252–258
- Zarowsky C, Haddad S, Nguyen VK (2013) Beyond ‘vulnerable groups’: contexts and dynamics of vulnerability. *Glob Health Promot* 20(1_suppl):3–9
- Zanetti LA, King CS (2013) Reflections on Theory in Action Transformational Public Service Revisited: Possibilities for Changing the Discipline?. *Administrative Theory & Praxis* 35(1):128–143

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.



Civic Vice in Digital Governance



Wessel Reijers

Abstract It has become a new global trend that governments are partially automating decision-making processes by public agencies. This, however, has led to some scandals revealing grave injustices, including the Robodebt scandal in Australia and the childcare benefit scandal in the Netherlands. This chapter argues that the normative impacts of the move towards automated decision-making can be fruitfully understood and addressed through the lens of civic virtue. It starts by outlining the Dutch childcare benefit scandal, showing that what happened cannot be reduced solely to human intent or structural factors, but needs to address the in-between term of human moral dispositions. Following this insight, the chapter outlines a framework of civic virtue, which outlines ideal states (civic virtue) and their deviations (civic vice) for different temporal configurations (past-, present-, and future-oriented civic virtue). Finally, the chapter uses this framework to reflect on the much-touted principle of ‘explainability’ in addressing harms like the ones done to citizens in the childcare benefit scandal. Three impacts are laid bare through the lens of civic virtue, of servility, presumptuousness, and political recalcitrance. Explainability, the chapter argues, may successfully address these impacts, but only if it ceases to focus solely on narrow black box problems in AI and address public governance more holistically.

1 Introduction

Since the rise of modern bureaucracies, there has been a growing concern about the consequences of introducing bureaucratic technologies and techniques into the everyday lives of citizens. The classical treatment of this problem was developed by Kafka in his *The Trial*, which depicts the arrest and trial of an innocent man, unable to understand or challenge his sentence. More recently, dystopian fictions like the movie *Brazil* depict how malfunctioning technologies that are ubiquitously applied

W. Reijers (✉)

Institute of Media Studies, Paderborn University, Paderborn, Germany

e-mail: wessel.reijers@uni-paderborn.de

© The Author(s) 2025

J. Goossens et al. (eds.), *Public Governance and Emerging Technologies*,
https://doi.org/10.1007/978-3-031-84748-6_14

315

in bureaucratic processes generate significant violations. In this case, technologies are used to sift between ‘good’ and ‘bad’ citizens, picking out any potential terrorists. Unfortunately, these stories are not limited to speculative fiction. Recent years have witnessed several serious scandals involving automated decision-making systems in benefit fraud detection. Cases like the Robodebt scandal in Australia and the childcare benefit scandal in the Netherlands show that severe rights violations can happen in relatively open and democratic societies. It is tempting to attribute these either to ill-willing individuals (e.g., racist civil servants) or to a technology-induced responsibility gap. Yet, neither of these explanations seems fully satisfactory. On the one hand, human, moral dispositions do matter, yet, on the other hand, these are always technologically and institutionally mediated. How to account for this situation?

A promising avenue to tackle this ethical conundrum is the surging literature on virtue ethics of technology, spearheaded by Shannon Vallor’s *Technology and the Virtues*.¹ On this count, we urgently need a strategy to cultivate the right technological virtues, ones that resonate with the demands and conditions of our technological society. This approach depends heavily on a conception of virtue anchored in individual flourishing, *eudaimonia*. The virtues and vices at stake, however, do not seem reducible to individual flourishing. Consider, for instance, a Dutch public servant who acts in a presumptuous (in Dutch: ‘vooringenomen’) manner while deciding on pre-labelled cases of suspected tax fraud. This disposition is in part motivated by a structure of public values and expectations, which finds its way to individuals through technologically mediated institutions. In other words, these cases concern dispositions that are situated *between* individuals and their political and institutional structures. They are about *civic* virtue and vice. Civic virtues and vices are those dispositions that citizens cultivate to support (or obstruct) the well-functioning of the (good) political community and its institutions. While resisting the reductionist impulse in theories of civic virtue, the concept seems to appeal to a plural range of dispositions, from Pettit’s notion of civility² to Arendt’s notion of anticipatory political action.³ Each of these conceptions puts demands on both the citizen and on the technologically mediated institutional environment in which he or she operates. As a guiding concept, civic virtue establishes both an ideal to strive towards and a deviation from this ideal to avoid; offering a basis to consider both virtue and vice in technologically mediated, institutional settings.

This chapter will focus on the emergence of civic vice through the use of automated decision-making in digital governance processes. In theoretical and practical efforts to realize responsible innovation of these systems, there is an increasing emphasis on ‘by design’ approaches that realize ethical principles like privacy and fairness. Of these principles, notably ‘explainability’ resonates with the core problem of automated decision-making systems, as the proclaimed opacity of these

¹ See Vallor (2016).

² Pettit (1997), p. 245.

³ Arendt (1958), p. 35.

systems makes it hard to include them in processes of public accountability and deliberation. As will be shown, this principle of explainability can arguably best be understood as a way to promote practices that cultivate civic virtue. The first section will delve into the case of the childcare benefit scandal in the Netherlands. It will argue that to understand the harm done by risk assessment systems involved, as well as similar systems developed by the Dutch authorities, we need to focus on how these systems mediate moral dispositions. The second section will address moral dispositions first through the lens of virtue ethics of technology, which, however, does not pay sufficient attention to the political dimension. To remediate this lack, a framework of civic virtue and civic vice is presented, as civic virtue is situated between ethics and politics. The third section reflects on the value of the much-touted principle of 'explainability' to address concerns raised by cases like the Dutch childcare benefit scandal. Evaluating the explainability of the systems involved through the lens of civic virtue, it shows how these systems contributed to the cultivation of civic vices of servility, presumptuousness, and political recalcitrance.

2 Digital Pride and Presumptuousness

In recent years, a series of serious government crises in Western liberal democracies has emerged out of a relatively new phenomenon: the increasing (partial) automation of digital government services that are central to the functioning of the welfare state. Quite infamously, the UK government decided during the COVID-19 pandemic to implement a rather simple algorithm to assign grades to students who were unable to achieve their A-levels.⁴ The allegedly unfair outcomes of this grading process, in which private schools enjoyed a disproportional advantage, led to a national outcry and British students protesting while exclaiming 'F*ck the algorithm.' In Australia, a 'policy fiasco' resulted from the application of a new, automated system to retrieve overpayments of social security, called 'Robodebt.'⁵ After multiple complaints about and investigations into this system, the government initiated a program to pay back over \$1 billion to over 400,000 citizens. The scandal led to the largest class action settlement in the history of the country. Another infamous scandal involving a digital system that partially automated the decision-making process happened in the Netherlands: the childcare benefit scandal. In what follows, the chapter will dissect this latter case more extensively.

In one of the greatest post-war miscarriages of justice in the Netherlands, thousands of parents and children were duped by the tax authorities. Parents were labeled as 'serious frauds' and had to pay back all the benefits they had gotten from the state

⁴See Ofqual (2020).

⁵See Whiteford (2021).

to organize otherwise expensive childcare.⁶ This often amounted to thousands or even tens of thousands of Euros, which they had to pay back within less than two years. Because of the economically fragile situation of many of these parents, the financial stress caused widespread misery. People lost their jobs and houses, got divorced, had their children placed out of their homes, and got serious physical and mental issues. Eventually, the scandal led to a government crisis, after which the government of prime minister Rutte felt compelled to step down, in January 2021.

A multitude of factors led to this tragic course of events. In an attempt to fight a rising number of cases of welfare fraud, the Dutch parliament had tasked the government with pursuing an extremely harsh fraud management.⁷ The tax authorities had taken on the task by issuing new measures based on a ‘business case’ that had to result in more money flowing from citizens back to the state. This included an all-or-nothing approach: once someone was labeled a fraud, he or she would get the label ‘intention/serious fault,’ which immediately implied that all received benefits had to be paid back. Being too understaffed to keep up with the increasing demand for scrutiny, the tax authorities implemented a new digital system for the evaluation of citizens, called *Fraud Signaling Facility* (FSV). The FSV automated the collection and evaluation of fraud signals, leading to a risk profile of groups of citizens.⁸

The system would gather large amounts of internal data (e.g., from tax declarations, benefit applications, citizen service numbers), and external data (e.g., from ‘tips and hints’ of fraud sent to an anonymous hotline).⁹ A risk model would then be used to predict, on the basis of these data, whether an application for benefits is likely to be faulty.¹⁰ Tax officers would still manually investigate whether ‘high risk’ for fraud indicated by the system was justified. If yes, they would tick a ‘fraud’ check box, which automatically put citizens on the trajectory of having to pay back thousands of Euros. External auditors had found that this check box was often ticked without the investigation having been completed, due to a lack of workforce capacity. When citizens would end up on a blacklist for frauds, there would be no systematic controls to verify the validity of the listings.¹¹ Even worse, there was no clear path for a citizen to get off such a list, for the very existence of this list was unknown to them.

The digital risk model used in the FSV was not the sole culprit in this affair. Members of parliament, who pushed for harsh fraud measures, ministers, who ignored signs of trouble, and civil servants, who pushed for the automated, harsh assessment, shared responsibility. In addition, there have been various cases of intentional discrimination based on nationality, which reflected the fact that most of the victims of the affair had a migrant background. And yet, as Jesse Frederik

⁶ See Tweede Kamer (2020).

⁷ See Frederik (2021).

⁸ KPMG (2020), p. 46.

⁹ See Autoriteit Persoonsgegevens (2021).

¹⁰ KPMG (2020), p. 51.

¹¹ Tweede Kamer (2020), p. 80.

explains, ‘civil servants in the tax department are tiny wheels in a largely automated process. They put a personal file next to a checklist to decide whether parents have a right to benefit or not.’¹² This argument is echoed by the findings of the parliamentary committee, publishing a report titled *Unprecedented Injustice*. ‘Tax officers did not consider it very likely that parents would have acted in good faith and that their applications could in fact be correct.’¹³ Trusting their risk models, tax officers had applied an arbitrary ‘80/20’ rule, assuming 80% of investigations would lead to the rightful confirmation of fraud. Clearly, the risk model generated the *impression* of the objective truth of risk indications offered by the system.

The FSV had not been the only attempt by the Dutch government to partly automate some of its processes of risk management. Another system, which was eventually only tried out in pilot projects, was the so-called *System for Risk Indication* (SyRI). SyRI is especially of interest because the system led to a court case, after which it had to be decommissioned. Developed in 2012, SyRI used ‘a certain combination of scores on indicators’ that ‘leads the system to identify persons (or organizations) with a higher risk of committing fraud.’¹⁴ In 2020, a group of six NGOs and two citizens sued the Dutch State, under the name of ‘Suspected in Advance.’ During the legal proceedings, the state chose not to disclose its risk model. It stated that secrecy had to be maintained, for otherwise, citizens would adjust their behavior to the risk model, i.e., ‘game the system.’¹⁵ The court, therefore, considered it impossible to judge whether SyRI really made ‘decisions’ and whether potentially a ‘self-learning algorithm’ might have been used. Yet, the laws that established SyRI as a legitimate technology, explicitly allow for this possibility. What is known is that SyRI *probably* had collected data on many aspects of people’s lives, including job status, real estate, data on housing, identity data, integration data (for migrants), history of compliance with the law, health insurance data, and many other categories. An ‘information bureau’ would subsequently pseudonymize the dataset and analyze it by applying the risk model. What would come out of the system is a blacklist of citizens, indicating the risk that they might be fraudulent.¹⁶ In its landmark ruling on February 5th, 2020, the Court judged SyRI unconstitutional, because it had grossly violated citizens’ fundamental right to privacy and because it had been kept secret.

The childcare benefit scandal led to a significant public response in the Netherlands. In this response, two ‘usual suspects’ in the discourse on technology ethics emerged, between what Latour aptly labeled the ‘materialist’ and the ‘sociological’ positions.¹⁷ The sociological position, on the one hand, focused on the

¹² Frederik (2021), p. 274.

¹³ Tweede Kamer (2020), p. 7. *Author’s translation*.

¹⁴ Bekker (2021), p. 294.

¹⁵ NJCM et al. v The Dutch State (2020) The Hague District Court ECLI: NL: RBDHA:2020:1878 (SyRI). Available at: <https://uitspraken.rechtspraak.nl/inziendocument?id=4ECLI:NL:RBDHA:2020:1878> [6.49].

¹⁶ See van Bekkum and Borgesius (2021).

¹⁷ See Latour (1994).

culture of institutional racism in the Dutch tax authority.¹⁸ Most of the victims of the childcare benefit scandal had a migration background, and tax officers had illegitimately acted against their best interest, by neglecting warning signals and even taking (foreign) nationality into account when deciding on a fraud case. The materialist position, on the other hand, emphasized technology. This position was quite clearly taken up by the activist group ‘Suspected in Advance,’ which positioned the SyRI system as a citizen sorting machine. Similarly, Frederik’s account is one of a tax authority in which individuals are mere cogs in a machine. Yet, while indeed both the moral stance of the tax authority and the involved civil servants and the ‘responsibility gap’ induced by the risk models are important pieces of the puzzle, neither side is sufficient by itself to explain what had happened. Instead, as Latour argues, a new ‘hybrid’ may have emerged out of the tax authority having started to use the new digital systems to manage risk.

To investigate this situation, I will focus on one of the main conclusions of the Parliamentary Committee that investigated the childcare benefit scandal, namely the argument that the tax authority had acted in a ‘presumptuous’ manner in handling suspected fraud cases.¹⁹ The tax authority and its civil servants were considered to have cultivated a negative *disposition*, of being presumptuous. The implicit idea that moral dispositions play an important role in the evaluation of automated systems and their outcomes has already been explored by scholars like Vredenburg. She argues that as (semi-)automated systems may act as street-level bureaucrats, we may evaluate their use in accordance with moral dispositions expected from civil servants (e.g., ‘the empathic elicitation of sensitive information’).²⁰ What is more, Vredenburg argues that the specific configuration of moral dispositions that is shaped by the introduction of automated systems may lead to ‘organizational pathologies.’ In other words, a relation persists between the moral dispositions, otherwise called virtues, of agents who played a role in the childcare benefit scandal, and the institutional structures in which those agents operate.

3 Civic Virtue and Vice

The previous section suggests that digital systems like the FSV and SyRI implicitly focus on establishing and maintaining a relation between the dispositions of citizens and (representatives of) political institutions. Perhaps we could in this context even speak of ‘collective virtues’, as argued by Astola et al.²¹ As the above discussion shows, the way that this relation is configured may go terribly wrong and cause significant harm. How, then, could we normatively argue about the modalities of

¹⁸ See Bot (2022).

¹⁹ Tweede Kamer (2020), p. 19.

²⁰ Vredenburg (2023), p. 22.

²¹ See Astola et al. (2022).

this relation? As we are concerned with (human) dispositions, virtue ethics appears to be a promising candidate for providing this normative perspective.

Virtue ethics has recently become a prominent approach in discourses on the ethics of emerging technologies, especially since Shannon Vallor's intervention in *Technology and the Virtues*. As an approach in normative ethics, virtue ethics is primarily interested in how human moral dispositions either contribute (in the case of virtue) or frustrate (in the case of incontinence or vice) the actualization of the 'being-completed' of human beings, which in Aristotelian virtue ethics is captured by the term *eudaimonia*. Vallor's book emphasized that apart from discussing the ethical impacts of technologies, for instance from a consequentialist point of view, we need to account for the way that technological change in the twenty-first century affects the virtues and vices. As such, technologies do not only produce benefits and harms, but may transform the characters of developers, policymakers, and users, for better or worse. Accordingly, the aim of virtue ethics of technology is to stipulate a strategy for cultivating our moral characters such that we adapt to and even flourish in the new digital age.²²

The virtue ethics approach has informed myriad of discussions in ethics of technology, some of which focus on particular technologies like AI,²³ while others focus on the application of virtue ethics in responsible innovation contexts.²⁴ Yet, it has also faced criticisms. It shares a shortcoming with the general field of technology ethics, in that it insufficiently considers the political impacts of emerging technologies. As scholars like Himmelreich²⁵ have demonstrated, technology ethics needs to engage with political philosophy in order to address the political dimensions of issues like fairness in AI systems. This lack has been raised more specifically in relation to the emergence of virtue ethics of technology. As Howard argues,²⁶ Vallor's approach focuses too much on the individual level without taking 'civic' technomoral virtues into account. In a way, Winner already anticipated this problem, arguing that citizens' virtues ought to be at the front of our minds when considering the impacts of new technologies.²⁷ Civic virtues, those dispositions cultivated by citizens to support the functioning of their political communities, have a decidedly political aspect that calls for our attention.

While for some discussions in technology ethics this lack of a political dimension may be unproblematic, for instance when discussing the extent to which care robots can show genuine care towards patients, for our purposes it represents a challenge. That is, systems like SyRI are clearly embedded in larger political institutions, which put political concerns at the forefront. Indeed, the harms that people experience or inflict are not merely related to individual ethics but also to their role

²² Vallor (2016), p. 6.

²³ See Constantinescu et al. (2021).

²⁴ See Tempels et al. (2017).

²⁵ Himmelreich (2020), pp. 33–35.

²⁶ See Howard (2017).

²⁷ See Winner (1995).

as *citizens* in relation to the institutions that impose these harms. For this reason, the dispositions that we are interested in, notably the ‘presumptuousness’ of civil servants, are not mere individual vices but are linked to ‘institutional pathologies.’²⁸

The core of the problem lies with the orientation of virtue ethics (and other ethical theories, including duty ethics and utilitarianism) towards the individual as the ultimate referent of meaning of the end or *telos* of the individual virtues. The meaning of courage, for instance, is derived from its relation to *eudaimonia*, i.e., an agent is courageous insofar as her disposition leads to the realisation of *eudaimonia*. This view, however, puts too much weight on the agent as being individually able to cultivate those ‘technomoral’ virtues to adequately adapt to the emerging technologies she is confronted with in the twenty-first century. As studies in moral psychology have shown,²⁹ this idea is naïve because dispositions of character are only stable in relation to a ‘situation,’ or a structure by which they are mediated.³⁰ This structure can be a physical one, for instance the architecture of a building, but also an institutional one, for instance in the case of a political system. In any instance, structures that mediate virtue are *technological*, which does not only mean involving technical artefacts but also technical skills, practices, and systems. This does not mean that political systems are somehow technological through and through, but that, as Wolff puts it, the actions they afford always have a technical aspect.³¹

This basic insight is not new but was already—even if not fully explicit—spelled out by Aristotle in *The Politics*. In that work, Aristotle distances himself to a certain extent from his account of virtue in the *Nicomachean Ethics*. Political or civic virtue, for Aristotle, is not simply the virtue of the good human being. Instead, it is lodged in between ethical virtue and political structures. As Aristotle argues, while the virtue of the good man is grounded in the individual *telos* of *eudaimonia*, civic virtue is relative to the constitution.³² As there are different constitutions and therefore different arrangements of political institutions in the Greek city states, there can be no ‘single and perfect virtue’ of civic virtue. Yet, the good political community should strive to make the virtue of the good man and the good citizen coincide, i.e., enable the good life for its citizens, which includes being able to co-decide in political matters. Hence, civic virtue is also not divorced from ethical virtue. Instead, as a concept, it *mediates* between ethical virtue and political institutions.

As civic virtue bridges the gap between ethics and politics that plagues virtue ethics of technology, it is a promising concept for arguing about the normative impacts of systems like FSV and SyRI. Elsewhere,³³ I have constructed a pluralist framework of civic virtue that draws from a plural range of theories of civic virtue (neo-republican, classical republican, and existential republican), conceptualized in

²⁸ Vredenburg (2023), p. 19.

²⁹ See Merritt et al. (2010).

³⁰ Reijers and Coeckelbergh (2020), p. 124.

³¹ Wolff (2021), p. 86.

³² See Pol. 3.4 1276b31–36.

³³ See Reijers (2023).

accordance with their temporal orientation. This temporal dimension was chosen because, as will become clear, these different theories of civic virtue distinguish themselves precisely by means of different conceptions of the temporality of human dispositions. In short, this framework distinguishes between three modalities of civic virtue that are important to consider in the light of emerging technologies that target citizen's dispositions: past-, present-, and future-oriented civic virtue.

First, past-oriented civic virtue denotes a disposition of 'civility' that emerges from reciprocal reputational rewards and sanctions exchanged between citizens.³⁴ It is past-oriented because it relies on a civic history of reputational events that accrue to a citizen's esteem. This notion is largely developed by the neo-republican tradition, for which civic virtue is instrumental in supporting institutions that promote freedom as non-domination.³⁵ Citizens are interested in being perceived benevolent in the eyes of others and engage in a distributed game of tit-for-tat social rewarding and sanctioning to maintain their economy of esteem. This activity is supported by what Pettit calls the 'intangible hand,' which is both distributed and operates non-intentionally (i.e., citizens manage their reputation without intending to). The intangible hand relies on techniques and technologies of organization, including those needed to organize the functioning of juries (e.g., jury rooms) and commons management (e.g., polycentric governance systems). In larger scale societies, where people do not directly or indirectly know one another, technologies like credit scores have become more pronounced to facilitate reputation management.³⁶ However, such technologies can also corrupt past-oriented civility. They may centralize the process of reputation management, for instance by having it managed by centralized governments or technology companies. Furthermore, they may turn reputation management into an intentional activity, meaning that citizens perform good deeds explicitly to gain reputation. As this affects the nature of the intangible hand, such new configurations may lead to the pathological structure of *domination*, in which agents are no longer to make their own choices in cultivating their disposition of civility from which they draw esteem. For instance, someone with a negative credit score might do things to gain reputation that are not at all related to their own conception of esteem, including 'unliking' certain types of music on social media. Under such circumstances, agents cultivate the problematic disposition or civic vice of servility.

Second, present-oriented civic virtue denotes a moral disposition aimed at realizing goods that are internal to a particular practice,³⁷ for instance, a conviction in political deliberation. This disposition is present-oriented because practitioners adhere to standards of excellence while acting in the appropriate manner in the present moment. Building on the classical republican tradition that stretches back to Aristotle, this notion considers civic virtue to be intrinsically valuable, requiring

³⁴ Brennan and Pettit (2005), p. 245.

³⁵ Pettit (1997), p. 245.

³⁶ Lauer (2017), p. 127.

³⁷ MacIntyre (2007), p. 190.

ethical and political virtue to (ideally) coincide in political participation. Citizens are interested in cultivating their civic virtue because it gives intrinsic meaning to the political practices they engage in. For instance, engaging in public speaking may cultivate a civic virtue of political conviction. The practices in question are defined through standards of excellence, which in turn are embedded in narratives belonging to individuals and to the larger political community. These narratives require public structures and institutions to be maintained, including those for purposes of civic education and public commemoration. There are cases of technologies being explicitly deployed to cultivate a sense of present-oriented civic virtue, as for instance with the Chinese Social Credit System.³⁸ Yet, present-oriented civic virtue can also be corrupted through technological mediation. When a technological infrastructure is built around a sense of efficiency or sanctioning systems of ‘carrots and sticks,’ citizens may engage in cultivating their virtues aimed at external rather than internal goods. This may lead to a culture of *conformism*, in which agents fail to gain the right motivation to cultivate civic virtue. For instance, citizens in China who want to improve their social credit information may engage in community action without intrinsic motivation, which is crowded out in favor of external gains. In this way, agents may cultivate the problematic disposition of presumptuousness.

Third, future-oriented civic virtue denotes an anticipatory disposition to act and speak in concert with others in the public sphere.³⁹ Such a disposition is future-oriented because the agent takes initiative, the effects of which are inherently unpredictable. This conception is largely influenced by what may be coined ‘existential republicanism,’ by which I mean republican political theory developed through engagement with existentialist and phenomenological philosophy—of which Arendt’s account is arguably the most influential. Future-oriented civic virtue is a disposition that may be exercised to initiate a political decision amongst a group of equal citizens. Citizens are interested in cultivating this sense of civic virtue as it allows them to enjoy political freedom, which, as political animals living in the polis, they inherently strive towards. For instance, citizens and politicians may initiate the process to form a new constitution, as happened in Chile in 2022. Unlike the other forms of civic virtue, there are no established rules or standards that may define future-oriented civic virtue, as it is defined only through its exercise of initiating something new. Yet, while the exercise of civic virtue is fragile, it needs institutional support by means of the public sphere. The public sphere is constituted by the world of things, and thus also the world of technology, including buildings (e.g., the ancient agora, the parliament) and procedures (e.g., parliamentary rules). At the same time, the world of technology has the potential to corrupt future-oriented civic virtue. A technology-infused public sphere may interfere with the agency of political actors, introducing new activities with temporalities different from political action, which Arendt coined labor and work. Such modulations subsume political action under processes of production and consumption, which degrades the action

³⁸ See Orgad and Reijers (2022).

³⁹ Arendt (1958), p. 199.

and makes it effectively apolitical. For instance, when political activism is conducted through online platforms it may degrade into citizens exploiting the technology to gain political ground and engage in ‘clicktivism.’⁴⁰ The degradation of political action may, in turn, give rise to the civic vice of ‘political recalcitrance.’⁴¹

The relation between ideal forms (leading to civic virtue) and deviations (leading to civic vice) depicted above share the same structure as those in Aristotle’s *The Politics*, in which ideal states are captured as kingship, aristocracy, and polity, while deviations are tyranny, oligarchy, and democracy. These modalities (see Table 1) should not be understood as absolutes, but rather as benchmarks to consider the great variety of possible dispositions and institutional structures one may encounter. Notably, technology plays an inherent role in the mediation of ideals (in the form of e.g., the intangible hand or a public sphere) and deviations. That is, as these concepts denote the institutional structures needed for the cultivation of civic virtue, they include a concern for the techniques and technologies that mediate these structures, including those supporting commerce (e.g., financial systems), civic education (e.g., school systems) and politics (e.g., voting systems). For this reason, we can map the kind of technological structures that support the ideal forms of civic virtue and their deviations. Certain technological structures will be conducive to civic virtue, while others are liable to cultivate what may be denoted as *civic vice*.

We arrive at a place of understanding the relation between (moral) dispositions and technopolitical structures. There exists a reciprocal relation, as for instance the public sphere co-constitutes anticipatory political action (enabling civic virtue), while tangible and intentional structures for the cultivation of civility may degrade civic virtue to dispositions of domination (enabling civic vice). We should note, however, that each modality of civic virtue relates to a particular sphere of human activities, much akin to Arendt’s distinction between labor, work, and action.⁴² How to ontologically distinguish between these spheres is beyond the scope of this

Table 1 Modalities of civic virtue and their ideal states and deviations

Type of civic virtue and vice	Ideal disposition (civic virtue)	Ideal structure	Deviation (civic vice)	Deviating structure
Past-oriented	Gaining honor	Intangible and non-intentional	Domination, leading to servility	Tangible and intentional
Present-oriented	Realizing internal goods	Standards of excellence embedded in community narrative	Conformism, leading to presumptuousness	Instruction based on gain of external goods
Future-oriented	Initiating political action	The public sphere	A-political action, leading to political recalcitrance	Systems of production and consumption

⁴⁰ See Fourcade and Johns (2020).

⁴¹ Button (2016), p. 102.

⁴² Arendt (1958), p. ix.

chapter, but as a rule of thumb we may count that political action requires future-oriented civic virtue, actions requiring professional discretion (i.e., work) require present-oriented civic virtue, and actions in the sphere of everyday economic exchange (i.e., labor) require past-oriented civic virtue. What is most important to note, is that not each activity can be judged against the same (high) level of civic virtue. This is also important in considering the scope of civic virtue. One might argue, for instance, that the civic virtue required of a civil servant in his or her professional capacity is not the same as that of a citizen engaging in political action, including a member of parliament taking new legislative initiatives. Yet, the above-mentioned framework emphasizes the existence of different modes of civic virtue that may apply in certain context of practice. Taking the same example, even if a civil servant does not need to live up to the demanding notion of political action, he or she would need to cultivate a sense of civility (which, as Pettit argues through different examples, may be linked to a sense of professionalism) but also a disposition of present-oriented moral excellence in certain situations, for instance ones that call for administrative discretion.

4 Explainability and Civic Vice

Now I have outlined the framework of civic virtue and vice, let us return to the case of the FSV and SyRI. The use of digital systems for risk management by the Dutch state, as exemplified by the cases discussed above, illustrates the ethical and political impacts at stake when crucial government services are increasingly automated. They challenge established public values like privacy and transparency,⁴³ which also seems to be reflected in policy responses to new digital technologies as used by public agencies. Most notably, the draft of the EU Artificial Intelligence (AI) Act aims to prohibit ‘the placing on the market, putting into service or use of AI systems by public authorities or on their behalf for the evaluation or classification of the trustworthiness of natural persons over a certain period of time based on their social behavior or known or predicted personal or personality characteristics.’⁴⁴ The AI Act exemplifies a policy push to mitigate the risks of systems like SyRI, while at the same time remaining vague on the purposes that would be deemed problematic, for instance by disallowing ‘detrimental or unfavorable treatment’ only in case this would be ‘unjustified or disproportionate.’

One of the primary regulatory and technical instruments to deal with the emergence of AI-powered systems is the much-touted principle of *explainability*.⁴⁵ Propelled by a policy push (it has already been integrated into the GDPR) and myriad technical advancements, the notion of explainability draws from the intuitive

⁴³ See van Bekkum and Borgesius (2021).

⁴⁴ European Commission (2021), p. 43.

⁴⁵ See Vilone and Longo (2021).

challenge that arises with the use of ‘black box’ AI systems. That is, while AI systems have inputs and outputs that may be understandable to humans, the process that constitutes the throughput of these systems cannot be directly ‘read.’ Instead, the many layers of weighted artificial neurons do not by themselves give an adequate explanation of *why* certain inputs lead to certain outputs. Explainability thus puts forward a requirement that this lack should be mitigated, and adequate ‘explanations’ may be given to consumers and citizens when AI systems make high-stakes decisions.

At face value, explainability does not seem to be an issue in the cases of the FSV and SyRI. Although this information is kept secret by the Dutch government (which contributed to the court’s judgement in the SyRI case), it seems unlikely that sophisticated machine learning was used in the FSV. More likely, good old-fashioned AI (GOFAI) was used to power a risk assessment model to sort through a large amount of data. Yet, I argue that the black box problem in fact distracts from the actual normative problem of explainability, which is not about the lack of an explanation per se but about the cultivation of a set of problematic dispositions: of civic vices.

As critics like Robbins have argued,⁴⁶ explainability in the narrow sense (i.e., making black boxes explicable) largely misses the point, because it circumvents questions around—for instance—relevance of a decision to be explained. The act of explanation is not merely the provision of an artefact (e.g., an explanation of why an automated system acted in a certain way) but is a process that takes place in a larger context.⁴⁷ As such, it seems that rather than an isolated machine (an AI system) deciding, a decision is in fact made by a larger ecosystem in which machines and humans collaborate. In comparison, we could consider another revolutionary machine, the steam engine. When it was introduced by industrialists like James Watt, it had not yet become ‘explainable’ in the language of thermodynamics, even while it practically worked. The normative problems arising from the steam engine, however, could also not be ‘explained’ through the science of thermodynamics—even if this science *contributed* to its safety, hence making a positive impact. Without the context of, amongst others, the emerging capitalist system, the organization of the factory, and the decisions made through the hierarchy of a firm, normative issues around principles like safety could not be addressed by means of making the machine more explainable.

All this is to say that explainability, while arguably important, is not merely about making AI systems explicable in the narrow sense. At the same time, this should not make us throw out the baby with the bathwater. When considering the case of the FSV, eerie equivalences with older themes like the one in Kafka’s *The Trial*, which often features in explainability literature, and even in Marx’s texts on alienation intuitively appear. Yet, these have nothing to do with machine learning per se. Instead, they point to a larger problem of (bureaucratic) automation, meeting the corruption of human dispositions. Explainability, perhaps better captured as a

⁴⁶ See Robbins (2019).

⁴⁷ O’Hara (2020), pp. 1–7.

capacity to act in deliberation⁴⁸ or to engage in self-advocacy,⁴⁹ instead of capturing a ‘property’ of an AI system, seems to denote a relation between human dispositions and technological and institutional structures. In this light, GOFAI systems can very well generate a call for explainability, as can other (even non-algorithmic) technologies of automation.

Hence, the lens of civic virtue—as introduced above—promises some new insights into the problem of explainability raised by systems like the FSV and SyRI. As these systems mediate different practices, including explicitly political practices (e.g., representatives deciding on new policies), administrative practices (e.g., civil servants deciding on fraud cases), and citizen’s practices (e.g., citizens being (un)able to appeal to decisions made), they can mediate the different modalities of civic virtue introduced above. Let us review each of those and discuss how they can add to an understanding of explainability.

First, *domination* may come about when a system configures a relation of tangible and intentional decision-making. Citizens derive their self-esteem (understood as reputation) from a distributed activity of tit-for-tat reputation management. Reputational judgment may be negative, but this negativity can only lead to civic virtue if it is not rendered tangible (e.g., centralized in a decision-making system) and if it is not intentionally designed to render a reputational judgment, for instance sorting ‘good’ and ‘bad’ citizens. While not explicitly designed with this purpose in mind, a system like the FSV implicitly incorporates these problematic aspects. The FSV both made tangible, recorded judgments concerning the status of citizens and did this with an explicit reputational purpose, one that is however expressed in terms of risk. This led to a dynamic of domination, according to which citizens were unable to exercise their agency in realizing their self-esteem as citizens. Such a dynamic may cultivate a problematic disposition, a civic vice, of servility. This also comes to the fore in research done on the impact of the childcare benefit scandal on citizens, who describe that they isolated themselves from others, lost their trust in institutions, and lost hope in good outcomes.⁵⁰ Rather than this being about explicating the recommendations made by the risk management system (narrow explainability), the issue lies with the very architecture of a system like the FSV. As such, it affects the ‘right to explanation’ in the broader sense, captured by Vredenburg as the effective right to informed self-advocacy.⁵¹

Concerning *conformism*, this deviation may come about when a system mediates the pursuit of external rather than internal goods of relevant practices. Citizens, especially in their professional capacities, derive their motivation in acting virtuously from practices that promote the goods internal to professional practices. For instance, as discussed above, in the case of street level bureaucrats, an internal good of engaging with citizen’s may be to cultivate empathic elicitation. Yet, in the case

⁴⁸ See Jongepier and Keymolen (2022).

⁴⁹ See Vredenburg (2022).

⁵⁰ See Konaté and Pali (2023).

⁵¹ See Vredenburg (2022).

of a risk assessment model like the FSV, the standards are derived from a 'business case' that aims to maximize efficiency of a bureaucratic process. The goods aimed at in this development are financial and material, which as such do not contribute to the standards of excellence of being a good civil servant dealing with citizens' tax records. As professionals consequently draw motivation from those ulterior standards, they risk conforming to a set of practices that is not conducive to their civic virtues. Instead, they may lead to civil servants acting presumptuously, failing to consider the citizen's situation and instead following a verdict proposed by a risk assessment system. Also here, merely having an explanation of this verdict would not have helped, especially because it has become clear that civil servants kept engaging in the problematic practices while knowing these were faulty. Instead, it again does resonate with explainability in the broader sense, as the use of the FSV failed to contribute to the moral dispositions (or rather civic virtues) asked of civil servants while discursively interacting with citizens.

Concerning *a-political activity*, this deviation may come about when a system mediates citizens' capacity to take initiative in the public sphere by subsuming it under a form of management. Citizens, most notably in their capacity as political representatives (e.g., members of parliament), act virtuously when engaging in future-oriented action and speech. Yet, the capacity to act politically may be relegated to systems like the FSV, which also corresponds to a new legal reality that some have captured as algorithmic governmentality.⁵² Instead of the parliament deciding on categorical laws, implying rights and duties, that it can change according to new political realities, the realm of law is partially delegated to (partially) automated systems. These systems, however, are inherently past-regarding and lack the capacity to take political initiative. As such, political action is increasingly overshadowed by a-political activities that are co-performed by systems and political actors managing these systems. This has become quite clear in the Dutch context, as politicians have found themselves empowered to delegate capacities to a fraud detection system (to fight fraud) but incapable of undoing the harms done by this system, which in fact haunts the political system up until this day. Politicians thus cultivate a problematic disposition of political recalcitrance, according to which they 'sacrifice the political relation among moral equals and the possibilities of reciprocal justice through a hardened disposition of indifference to the valid claims and considerations of others.'⁵³ Here, again, explainability in the narrow sense would not have helped significantly to reinstate civic virtue. Nevertheless, explainability in a broader sense, aimed at restoring the core deliberative agency of political actors, seems to resonate strongly with the issue at hand.

One may object, at this point, that the concept of explainability has been inflated beyond its legitimate boundaries in the above discussion. Explainability, according to this view, should be limited to a formal-procedural criterion which exclusively

⁵² See Rouvroy and Berns (2013).

⁵³ Button (2016), p. 102.

applies to machine decisions (and mostly to AI).⁵⁴ I would not disagree with such a view if it would also downplay the importance of explainability in normative and policy discourses. Yet, given the current normative weight applied to explainability, it seems to me that broadening rather than narrowing the concept is the right approach. It does in this regard not matter too much what term or cluster of terms is used (explainability, explicability, interpretability, etc.) but rather what field of normative concerns is carved out by invoking it. Narrowing down the meaning of explainability by exclusively applying it to interpretations of outcomes of machine learning systems causes blind spots in e.g., policy discussions concerning normative problems that come with systems like SyRI. The question should not be “was machine learning used?” but rather “are relevant explainability concerns at stake?” Just as data protection is an issue for surveys performed with pen-and-paper, so explainability concerns constitute an issue beyond machine learning, including Kafka’s pen-and-paper bureaucracies. That is, these normative principles in decision-making do not lose their force once machine learning is not in the picture, as good-old-fashioned risk assessment systems may similarly trigger explainability concerns.

Moreover, it may be objected that the language of ‘rights’ and, relatedly, interests, would be more suited to cover the concerns discussed above, rather than the language of civic virtue. To this, I would reply that most discussions appealing to a right to explanation at some point in the argument are confronted with a criterion of *relevance*. That is, the right to explanation is only relevant in relation to certain practices (e.g., those invoking core deliberative capacities) but how are such practices to be delineated? An appeal to interests could do the trick here, but also interests are not easy to ground conceptually. As Arendt already argued, humans have interests because they live in a shared world. Yet, the question of what interests should be considered relevant and which not (i.e., which should have a ‘moral weight’) does not lend itself to a straightforward answer. The lens of civic virtue, as discussed above, is a potential answer to these lacunae, as it generates a direct link between technologically mediated practices and the relation between citizens’ dispositions and the larger institutions in which citizens act.

5 Conclusion

In this chapter, the lens of civic virtue was developed and used to investigate the normative impacts of the increasing (partial) automation of decision-making by public agencies. By means of this investigation, the chapter argued that this development brings the risks of (1) increasing domination, leading to servility, (2) increasing conformity, leading to presumptuousness, and (3) political action being transformed into non-political action, leading to political recalcitrance. The

⁵⁴ See Zödi (2022).

progressive cultivation of these civic vices may hollow out democracy and the rule of law, as it affects citizens in their everyday activities, in their professional lives, and in their attempts to participate in political affairs. While the newly proposed principle of ‘explainability’ may be useful in addressing these risks, it may only do so if it moves beyond narrow considerations of the black box problem in AI. Instead, explainability needs to be made explicit as a broad principle applying to public governance, one that brings out requirements for institutions and the technologies they enlist to foster and not diminish a sense of what may be coined ‘practical wisdom.’⁵⁵ This, however, is not merely a question of building better machines. Rather, design, policy making, and political activism will need to converge, both consensually and through contestation, to shape the requirements needed to avoid future tragedies like the Dutch childcare benefit scandal.

References

- Arendt H (1958) *The human condition*. University of Chicago Press
- Aristotle (1998) *Politics* (C. D. C. Reeve, Trans.). Hackett Publishing Company. (Original work published ca. 350 BCE)
- Astola M, Bombaerts G, Spahn A, Royakkers L (2022) Can creativity be a collective virtue? Insights for the ethics of innovation. *J Bus Ethics* 179(3):907–918. <https://doi.org/10.1007/s10551-021-04833-0>
- Autoriteit Persoonsgegevens (2021) *Onderzoeksrapport Belastingdienst Verwerkingen van persoonsgegevens in de Fraude Signalering Voorziening (FSV)*
- Bekker S (2021) Fundamental rights in digital welfare states: the case of SyRI in the Netherlands. In: Spijkers O, Werner WG, Wessel RA (eds) *Netherlands Yearbook of International Law* 2019, vol 50. Springer, pp 289–307
- Bot M (2022) Is institutioneel racisme echt racistisch? *Nederlands Juristenblad* 26:2048–2055
- Brennan G, Pettit P (2005) *The economy of esteem: an essay on civil and political society*. Oxford University Press
- Button ME (2016) *Political vices*. Oxford University Press
- Constantinescu M, Voinea C, Uszkai R, Vică C (2021) Understanding responsibility in Responsible AI. *Dianoetic virtues and the hard problem of context*. *Ethics Inf Technol* 23(4):803–814. <https://doi.org/10.1007/s10676-021-09616-9>
- European Commission (2021) *Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain Union legislative acts*, Pub. L. No. COM(2021) 206 final
- Fourcade M, Johns F (2020) Loops, ladders and links: the recursivity of social and machine learning. *Theory Soc* 49(5–6):803–832. <https://doi.org/10.1007/s11186-020-09409-x>
- Frederik J (2021) Zo hadden we het niet bedoeld: De tragedie achter de toeslagenaffaire. *De Correspondent*
- Himmelreich J (2020) Ethics of technology needs more political philosophy. In: Chien AA (ed) *Communications of the ACM: techniques for interpretable machine learning*. Association for Computing Machinery, pp 33–35. <https://doi.org/10.1145/3339905>
- Howard D (2017) Technomoral civic virtues: a critical appreciation of Shannon Vallor’s technology and the virtues. *Philos Technol* 31:293–304. <https://doi.org/10.1007/s13347-017-0283-1>

⁵⁵ See Karlan and Allen (2024).

- Jongepier F, Keymolen E (2022) Explanation and agency: exploring the normative-epistemic landscape of the “Right to Explanation”. *Ethics Inf Technol* 24(4):49. <https://doi.org/10.1007/s10676-022-09654-x>
- Karlan B, Allen C (2024) Engineered wisdom for learning machines. *J Exp Theor Artif Intell* 36(2):257–272. <https://doi.org/10.1080/0952813X.2022.2092559>
- Konaté S, Pali B (2023) “You have to talk with us, not about us”: exploring the harms of wrongful accusation on those affected in the case of the Dutch ‘childcare-benefit scandal’. *J Victimol* 16:139–164
- KPMG (2020) Rapportage verwerking van risicosignalen voor toezicht. Available at: <https://open.overheid.nl/documenten/ronl-42970d17-d8b9-41d6-aa8f-d0cc52ab97c8/pdf>
- Latour B (1994) On technical mediation—philosophy, sociology, genealogy. *Common Knowledge* 3(2):29–64
- Lauer J (2017) Creditworthy: a history of consumer surveillance and financial identity in America. Columbia University Press
- MacIntyre A (2007) After virtue: a study in moral theory. University of Notre Dame Press
- Merritt MW, Doris JM, Harman G (2010) Chapter 11: Character. In: Doris JM (ed) *The moral psychology handbook*. Oxford University Press, pp 355–401
- O’Hara K (2020) Explainable AI and the philosophy and practice of explanation. *Comput Law Secur Rev* 39:105474. <https://doi.org/10.1016/j.clsr.2020.105474>
- Ofqual (2020) Awarding GCSE, AS, A level, advanced extension awards and extended project qualifications in summer 2020—Interim report. Ofqual
- Orgad L, Reijers W (2022) How to make the perfect citizen? Lessons from China’s Model of Social Credit System. *Vanderbilt J Transnatl Law* 54(5):1087–1121. Available at SSRN. <https://doi.org/10.2139/ssrn.3586503>
- Pettit P (1997) Republicanism: a theory of freedom and government. Oxford University Press
- Reijers W (2023) Technology and civic virtue. *Philos Technol* 36(4):71. <https://doi.org/10.1007/s13347-023-00669-w>
- Reijers W, Coeckelbergh M (2020) Narrative and technology ethics. Palgrave Macmillan
- Robbins S (2019) A misdirected principle with a catch: explicability for AI. *Minds Mach* 29(4):495–514. <https://doi.org/10.1007/s11023-019-09509-3>
- Rouvroy A, Berns T (2013) Algorithmic governmentality and prospects of emancipation. Disparateness as a precondition for individuation through relationships? (L. Carey-Libbrecht, Trans.). *Rezeaux* 177(1):163–196. <https://doi.org/10.3917/res.177.0163>
- Tempels T, Blok V, Verweij M (2017) Understanding political responsibility in corporate citizenship: towards a shared responsibility for the common good. *J Glob Ethics* 13(1):90–108. <https://doi.org/10.1080/17449626.2017.1320577>
- Tweede Kamer (2020) 35 510 Parlementaire ondervraging Kinderopvangtoeslag. De Parlementaire Ondervragingscommissie Kinderopvangtoeslag
- Vallor S (2016) Technology and the virtues: a philosophical guide to a future worth wanting. Oxford University Press
- van Bekkum M, Borgesius FZ (2021) Digital welfare fraud detection and the Dutch SyRI judgment. *Eur J Soc Secur* 23(4):323–340. <https://doi.org/10.1177/13882627211031257>
- Vilone G, Longo L (2021) Notions of explainability and evaluation approaches for explainable artificial intelligence. *Inf Fusion* 76:89–106. <https://doi.org/10.1016/j.inffus.2021.05.009>
- Vredenburg K (2022) The right to explanation. *J Polit Philos* 30(2):209–229. <https://doi.org/10.1111/jopp.12262>
- Vredenburg K (2023) AI and bureaucratic discretion. *Inquiry*:1–30. <https://doi.org/10.1080/0020174X.2023.2261468>
- Whiteford P (2021) Debt by design: the anatomy of a social policy fiasco – or was it something worse? *Aust J Public Adm* 80(2):340–360. <https://doi.org/10.1111/1467-8500.12479>
- Winner L (1995) Citizen virtues in a technological order. In: Feenberg A, Hannay A (eds) *Technology and the politics of knowledge*. Indiana University Press, pp 65–84

- Wolff E (2021) Between daily routine and violent protest: interpreting the technicity of action. De Gruyter
- Zödi Z (2022) Algorithmic explainability and legal reasoning. *Theory Pract Legis* 10(1):67–92. <https://doi.org/10.1080/20508840.2022.2033945>

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.



Algorithmic Governance and Social Vulnerability: A Value Analysis of Equality and Trust



Antonia Stanojević

Abstract This chapter presents a value analysis of equality and trust using cautionary examples of algorithmic governance that threaten these public values. The chapter uses experimentation as a conceptual lens to address the antecedents and consequences of the premature implementation of algorithmic governance targeting disproportionately socially vulnerable individuals and communities. The use of algorithmic governance before enough is known about the potential ill effects is tantamount to experimentation, and the potential harms from being targeted for testing such systems are disproportionately consequential for vulnerable populations, in economic and public value terms alike. The chapter concludes with a policy recommendation that strongly advises against the experimentation with algorithmic governance on vulnerable populations, because this deteriorates equality and trust—values that are already comparatively deprived among the socially vulnerable members of society. Therefore, as a moral obligation, social vulnerability should be considered a counterindication for the nascent use of algorithmic governance, particularly in the shape it tends to take currently—that of surveillance and control.

Funding: This chapter was written within the project “The Role and Responsibilities of Public Actors in Distributed Networks. Transparency, Trust and Legitimacy by Design” [project no. MVI.19.040], financially supported by the Dutch Research Council (NWO) as part of the project within the framework of the program “Responsible Innovation. Designing for public values in a digital world.”

A. Stanojević (✉)
Tilburg Law School, Tilburg University, Tilburg, The Netherlands
e-mail: A.Stanojevic@tilburguniversity.edu

1 Introduction

In 2021, the Dutch government was forced to resign over what became globally known as the childcare benefits scandal, after a parliamentary inquiry into the matter concluded that fundamental principles of the rule of law had been violated.¹ The childcare benefits scandal has since then been used as a warning to governments that increasingly seek to algorithmize their systems, as it appears that the scandal was (at least in part) caused by the use of data driven decision-making. Specifically, a self-learning algorithm for fraud detection used as a part of a semi-automated system wrongly flagged citizens who were then asked to repay the benefits they were given and to pay the issued fines. Consequently, tens of thousands of families were driven into poverty. Many of the citizens harmed by this policy belonged to ethnic minorities, notably Turkish and Moroccan,² and an audit later showed that having a foreign nationality was one of the determining factors used for flagging by the algorithm.³ Two years after this scandal, the Dutch minister of education halted the use of a similar algorithm for student loan fraud detection, until more is known about ethnic discrimination suspected also in this case.⁴ Although the algorithm for student loan fraud detection did not make predictions based on ethnic background data, presumably in an active attempt to prevent direct discrimination, it appears it ran on enough proxy data to produce a biased effect with regard to ethnic background, since a whopping 97% of the appealed decisions were issued to students with a migration background.⁵

There are several issues surrounding these cases that warrant academic inquiry, most notably prejudice and discrimination based on ethnic background, which appear to be pervasive biases in both human-driven and data-driven decision-making. Next to the questions of bias and the associated public value of equality, these examples provoke other classical public value questions associated with emerging technologies, such as questions around trust. As can be deduced from the above examples of algorithmic fraud detection gone wrong, algorithmic governance has the potential to erode certain public values. I will address these effects on public values through the conceptual lens of *experimentation*.

The notion of emerging technology implementation as experimentation has been around for decades, ever since Krohn and Weyer's⁶ article was published in *Science and Public Policy*, in which they refer to societal implementation of technological innovations as 'real-life experiments' or 'implicit experiments' (p. 173). However, real-world experimentation is not limited to the domain of technology—if new technologies being released in the real world can be considered an implicit experiment,

¹ See Ondervragingscommissie Kinderopvangtoeslag (2020).

² See Heikkilä (2022).

³ See Ten Seldam and Brenninkmeijer (2021).

⁴ See Ersoy and Van der Gaag (2023).

⁵ See De Heer (2023).

⁶ See Krohn and Weyer (1994).

then so can new policies being released in the real world. Indeed, Dewey⁷ argued that introducing new policies can also be considered an experiment of sorts. In line with both of these arguments, i.e. the unknown technological aspects and the unknown societal implications, van de Poel⁸ advances the experimentation framework by conceptualizing the widespread implementation of emerging technologies as sociotechnical experimentation. In this chapter, I argue that algorithmic governance fits within the sociotechnical experimentation framework perfectly, as it is compatible with both arguments. Firstly, algorithms as technologies that are continuously developing based on the feedback resulting from their implementation satisfy the technological experimentation argument, and secondly, integrating them into novel policies with unknown and to a large extent unpredictable societal outcomes satisfies the social experimentation argument. As such, algorithmic governance can be considered a prime example of sociotechnical experimentation.

Within the experimentation framework, algorithmic governance can be conceptualized as a novel policy, one that is still under development, and must as such be “beta tested” on human subjects. However, even if we accept as a given the phenomenon of experimentation with algorithmic governance, questions remain around who should be targeted by experimental algorithmic governance policies. In particular, why are there so many examples, globally, of algorithmic governance being *de facto* tested on socially vulnerable individuals and communities, often with far-reaching negative consequences on these members of society?

For clarity, let us define social vulnerability as a characteristic that emerges from differential group relations in society⁹ and encompasses social, cultural, economic, political, and institutional processes that influence how one experiences hazards or adversities.¹⁰ It is important to note that this definition, while acknowledging the implications of social vulnerability on the individual’s experience of adversity, does not define vulnerability as an inherent characteristic of the individual—on the contrary, it emphasizes power relations that position certain groups at a disadvantage compared to others (*social* vulnerability). Moreover, the definition allows for intersectional vulnerability created by one’s position within overlapping systems of power.

The concept of social vulnerability overlaps to a great extent with the concepts of social, spatial, and economic marginality,¹¹ defined as exclusion from mainstream economic, educational, or cultural life.¹² In practice, it also overlaps to a great extent

⁷ See Dewey (1938).

⁸ See van de Poel (2016).

⁹ See Fordham et al. (2013).

¹⁰ See Spielman et al. (2020).

¹¹ Marginality is, of course, relative, or dependent on subjective (and often colonial) definitions of the center and periphery (Cullen and Pretes 2000). For the purpose of this chapter, marginality is defined in relation to the center defined as the hegemonic web of norms (including neoliberal market solutions, globalization and digitalization), and periphery as everything that deviates from this norm, as explained in main text.

¹² See Pelc (2017).

with lower socio-economic status, therefore making vulnerability often synonymous with poverty. Within the conceptual framework of experimentation, however, social vulnerability is the precise characteristic of interest, because experimentation, as will be shown throughout the chapter, represents a potential hazard in its own right, due to potential unforeseen negative consequences.

Since vulnerability (as opposed to resilience) defines the success of overcoming adversities, the concept of social vulnerability, rather than poverty or marginality, is the most appropriate within the theoretical framework of this chapter, while acknowledging its connection to various socio-economic indicators. Examples of social vulnerability include, but are not limited to: senior, disabled and homeless individuals, undocumented individuals, ethnic and racial minorities, or members of certain rural, remote or low-income communities. These individuals and groups are specifically considered in the context of algorithmic governance experimentation, as this chapter presents examples that indicate disproportionate experimentation on the socially vulnerable.

Besides social vulnerability, which emphasizes the societal power dynamics as sources of individual vulnerability, other vulnerability theories emphasize the universal,¹³ situational¹⁴ and contextual¹⁵ nature of vulnerability, as detailed in Chapter 12 of this volume (Ranchordás & Beck). These vulnerability theories are compatible—while we are all vulnerable as a feature of the human experience and embodiment (universal vulnerability), some of us are more vulnerable than others (social vulnerability). Moreover, it is plausible that situational and contextual vulnerability “lean on” social vulnerability, creating a synergistic effect of different kinds of vulnerability that differentially affects various social groups. However, this chapter maintains the focus on social vulnerability, as the analysis identifies exactly this trait as being targeted in the current examples of algorithmic governance experimentation, and simultaneously as the precise trait that should be considered a counterindication for algorithmic governance experimentation.

This chapter aims to analyze what happens when experimental modes of algorithmic governance, such as self-learning algorithms, are used to make decisions over vulnerable citizens. This practice does not seem advisable, since vulnerable citizens are by definition *a priori* vulnerable to hazards, i.e. to the potential unpredictable outcomes, mistakes and negative side effects of the novel algorithmic governance systems, such as having to pay incorrectly issued fines. Moreover, from an ethical standpoint, values such as equality and trust are upheld to a comparatively lesser degree among socially vulnerable individuals and communities, and experimentation with algorithmic governance further hinders both of these public values. It appears, thus, that experimentation with algorithmic governance on vulnerable citizens “makes the poor poorer” in both value and economic terms.

¹³ See Fineman (2008).

¹⁴ See Mackenzie (2014).

¹⁵ See Luna (2009).

In the aforementioned examples, policies relying on algorithmic governance were implemented with the aim to detect and prevent fraud, indicating low trust of the government in the communities on the receiving end of these policies (childcare benefits users; students). This chapter considers the bidirectional relational nature of trust, while approaching trust as a public value worth cultivating in society. The second public value salient in these examples is equality, diminished firstly through targeting vulnerable citizens with a self-learning algorithm (equality before the algorithm), and secondly through higher likelihood of being wrongly flagged due to ethnic background (equality after the algorithm). The chapter makes the case that trust and equality are historically deprived among vulnerable communities, and algorithmic governance experimentation poses a risk for further erosion of these values. The risk of extreme value deprivation is therefore disproportionate in case of algorithmic governance experimentation targeting vulnerable communities.

To summarize, this chapter brings forth a value analysis of algorithmic governance policy experimentation with emerging technologies that particularly concern socially vulnerable citizens. To this end, I begin by introducing experimentation as a conceptual lens; followed by considering how experimentation with algorithmic governance affects each of the relevant public values (equality and trust, respectively); and finally conclude with policy recommendations based on the conducted value analysis.

2 Algorithmic Governance as Policy Experimentation

The word experimentation has two meanings: experimentation as a scientific method, and experimentation as an exploratory approach.¹⁶ Although these two meanings are distinct, they are connected by a common denominator—namely, the uncertainty of the outcome, which is also the motivation behind experimentation. Experimentation is as such often used in public policy, where it can take on either of these meanings. Policy experimentation is defined as a flexible, meaning amendable and reversible, evidence-based approach to policy making.¹⁷ Both elements of this definition are necessary for the proper functioning of policy experimentation—while the evidence-based approach achieves policy learning, the reversibility of policy experiments safeguards the experiment subjects (the population) from unpredictable adverse effects of policy experiments.¹⁸ In the examples of algorithmic governance experimentation provided in the remainder of the chapter, it remains to be seen to what extent these instances of policy experimentation can be described as reversible and evidence-based.

¹⁶ See Huitema et al. (2018).

¹⁷ See Tassey (2014); McFadgen and Huitema (2018).

¹⁸ See Listokin (2008).

Policy experimentation more commonly entails an experimental approach to governance through a trial-and-error approach¹⁹ than experiment as a scientific method. Experiment as a scientific method in a strict sense refers to controlled randomized experiments, but in the domain of public policy, such experiments are often impossible for practical and ethical reasons. Therefore, in public policy, quasi-experiments (not randomized) and natural experiments (not randomized nor controlled) are more commonly used. Exploratory experimentation in public governance also has two variants: “Darwinian experimentalism”, focused on trying out diverging approaches with the aim to compare them, and “generative experimentation”, focused on trying out and improving one approach.²⁰ The latter, generative experimentation, seems particularly relevant for algorithmic governance, whereby it seems that algorithmic governance is accepted by policymakers as a path towards good governance.²¹ As such, algorithmic governance experimentation most commonly seems to take the generative route, whereby it serves to perfect the already established direction for policy development—one of increasing digitalization and algorithmization of governance.

Policy experimentation is, therefore, known to occur in three forms, namely (from least to most common) as: scientific experiment (with differing degrees of rigor), Darwinian experimentalism, and generative experimentation. Algorithmic governance experimentation can theoretically take any of the above forms, which differ with regard to the motivation (to test a hypothesis, compare diverging approaches, or improve a developing approach) and the level of methodological rigor surrounding policy experimentation. However, what positions most cases of algorithmic governance as experimental is their exploratory character, novelty, and unpredictability. This is certainly the case from the citizens’ perspective, for whom the motivation and the degree of methodological rigor are opaque. Rather, the citizens’ perspective is informed by the lived experience of human subjects who are on the receiving end of emerging algorithmic governance policies, especially when such policies result in mistakes and negative outcomes, as in the childcare benefits scandal. In such cases, failing to prevent mistakes and mitigate negative outcomes when mistakes do occur, indicate a lack of knowledge regarding the policy impacts before implementation, which signals their experimental character to citizens.

The ethical question then, underpinning this chapter, is: how does the described experimental approach to testing and/or developing algorithmic governance policies affect the socially vulnerable, who seem so often to be its target? To this end, let us first consider theoretical support, followed by practical examples, for the notion that the socially vulnerable are indeed commonly the target of experimentation, including experimentation with algorithmic governance. Historically, social power has undeniably been the factor determining one’s likelihood of being a test

¹⁹ See Collingridge (1992).

²⁰ See Ansell and Bartenberger (2016).

²¹ See European Commission (2024).

subject in the development of ever-emerging technologies,²² most of them throughout the past medical,²³ but these days increasingly digital.²⁴ As Bikundo²⁵ notes, discrimination always was, and still is, a core element of experimentation, as the Foucauldian power-knowledge axis remains the divide between the experimenters and experiment subjects. In digitalization literature, this power asymmetry is termed the “big data divide”, denoting the division between those who collect data and those from whom the data is collected.²⁶

This notion is substantiated and expanded in Virginia Eubanks’ seminal work *Automating Inequality: How High-Tech Tools Profile, Police and Punish the Poor*,²⁷ in which she argues that algorithmic governance exacerbates existing inequalities and privileges. In particular, its targeted use on poor and working-class people represents merely an evolution of the existing tendencies to police and punish the poor, thereby reinforcing the status quo. Perhaps this tendency to police the poor contributed to the fact that historically more data was collected on poor people, which continues to be the case today.²⁸ As such, data availability and societal power disparities intertwine and synergistically contribute to the observed trend of algorithmic governance experimentation targeting vulnerable communities. The specific examples from around the world, given through value analysis in the following sections, shed light on the reasons behind the tendency to disproportionately surveil and control the poor, or in the language better suited to the conceptual framework of experimentation adopted in this chapter, the vulnerable members of society.

3 Equality

Referring back to the definition of social vulnerability as relatively lower resilience to hazards shaped by group differences, it is clear that equality as a value is inherently deprived among vulnerable groups and individuals. Value analysis of equality is therefore of key importance, because any potential biases present in experimental implementations of algorithmic governance that result in inequality compound the already eroded public value of equality among the targeted vulnerable populations. When discussing how equality is affected by the use of algorithmic governance, one must first address the elephant in the room, which is that algorithms are trained on data, and data is necessarily a flawed representation of reality.²⁹ Aside from the

²² See Kopp (2006); Negri (2017).

²³ See Appleman (2020).

²⁴ See Taylor (2021); van Zoonen (2020); Sandvik et al. (2017); Metcalf and Crawford (2016).

²⁵ See Bikundo (2020).

²⁶ See Andrejevic (2014).

²⁷ See Eubanks (2018).

²⁸ See Madden et al. (2017).

²⁹ See Rodriguez (2022).

typical data issues, such as biased samples and measurement errors, the problem with data used for algorithmic governance is that it often contains data entries reflective of previous decisions made by humans, and humans are infamous for our wide array of cognitive biases, especially when it comes to social cognition. In computer science, the notable maxim goes “garbage in, garbage out”, meaning that if the data is flawed, the output will by necessity also be flawed. Considering that many human decisions are influenced by deeply entrenched biases, such as ethnic and racial biases, aggregating those decisions to guide future decisions must be done with caution and awareness of systemic biases, such as societal stereotypes and implicit and institutional racism.

The issue of data reflecting potentially biased human decisions becomes particularly evident when algorithms are applied in public governance, because biased algorithmic outputs informing public governance are exceedingly consequential. For example, an algorithm called COMPAS was used in the U.S. court system to calculate risk of re-offence, based on data concerning previous offences,³⁰ which are obviously biased due to previous human decisions, such as historical over-policing of the African American community and racially biased judgements. Both the data this algorithm was trained on, and the data it actually analyzed, likely contained these biases. Lo and behold, the algorithm produced high risk scores two and a half times more often for black than for white offenders, and these scores were twice as likely to be wrong, i.e. black offenders were almost twice as likely to be labeled high risk, but in reality, not re-offend.³¹ This exemplifies an experimental algorithmic governance policy targeting vulnerable US citizens (ex-convicts), and it resulted in inaccurate, racially biased decisions. Similarly, in the Dutch childcare benefits example, vulnerable citizens who relied on childcare benefits were targeted by the algorithmic governance policy, which also resulted in racially and ethnically biased decisions. These examples demonstrate that algorithmic governance poses risks for the societal value of equality on at least two levels—equality with regards to targeting the governed populations (equality before the algorithm—*ex ante equality*), and equality with regards to non-discrimination by the algorithm (equality after the algorithm—*ex post equality*).

Equality before the algorithm could be achieved by equalizing the likelihood across groups, individuals, and organizations to become test subjects in experimental implementations of algorithmic governance. This solution seems far more conducive to the public value of equality than disproportionately targeting vulnerable populations with algorithmic governance, which seems to be the case at present. On the other hand, promoting equality as a public value could also be achieved by applying the principle of equity to the selection of test subjects, by prioritizing more resilient actors for algorithmic governance experimentation, such as for example big corporations. Similarly, to achieve equality after the algorithm, two conflicting approaches can be distinguished, namely *equality of treatment (blindness) approach*

³⁰ See Minow (2023).

³¹ See Angwin et al. (2016).

and *equality of outcomes (awareness) approach*. These approaches differ with regards to how they treat protected characteristics associated with historical discrimination, such as race and gender—the former treats everyone equally, regardless of the protected characteristics (i.e. the color-blind approach) and the latter instead acknowledges the protected characteristics and adjusts the outcomes accordingly (e.g. gender quotas).

To ground this logic in the algorithmic landscape, I follow a recently published analysis of a hypothetical discrimination lawsuit against Facebook,³² and extrapolate this analysis to the Dutch fraud detection examples. The hypothetical example concerns Facebooks' algorithm that predicts the likelihood that a person will click on a certain job advertisement. In this case, the problem is that the algorithm, accurately based on the input data, shows lower paying vacancies to women, because women are more likely to click on those ads. The equal treatment (blindness) approach would entail taking gender as a variable out of the predictive algorithm. However, the issue here is that this does not make the algorithm "blind" to gender, because gender can still be determined from proxy variables that correlate with it. In fact, in this hypothetical example, the algorithm produces greater gender disparities when gender is not explicitly listed as a predictor variable. This phenomenon is not only hypothetical, however, as experiments found that, in cases where one group is already disadvantaged by algorithmic predictions, simplification by reducing the number of variables in the algorithm will necessarily further hinder fairness and equity.³³ It is difficult to overlook the parallel of the hypothetical Facebook example with the Dutch fraud detection examples, particularly when it comes to the effects of algorithm simplification by removing protected characteristic. Namely, in the childcare benefits scandal, the protected characteristic of ethnicity was used in the predictive algorithm, which resulted in inaccurate, ethnically biased results. Although the inquiry into the student fraud detection algorithm that followed is still ongoing, it seems that authorities wanted to prevent repeating the same mistake by opting for exclusion of ethnic background as a predictor variable, which still resulted in ethnic bias.³⁴

The equal outcome (awareness) approach would entail constraints of the outcome by employing so-called Disparate Learning Processes (DLPs), such as, in the case of the hypothetical Facebook example, equalizing the average pay between job listings shown to men and women. However, Simons and colleagues³⁵ argue that this would reduce the accuracy by showing both men and women job listings they are not interested in. Moreover, Lipton and colleagues³⁶ show that constraints of DLPs can produce unpredictable new biases, such as in this example showing higher paid jobs to more "manly" women. So, is there a way to apply the DLP constraints

³² See Simons et al. (2021).

³³ See Kleinberg and Mullainathan (2019).

³⁴ See Hoger Onderwijs Persbureau (2024).

³⁵ See Simons et al. (2021).

³⁶ See Lipton et al. (2018).

to the childcare benefits and student loan fraud detection examples, and what would happen in that case? One possible solution would be constraining the number of ethnic minority fraud cases according to their proportion in the population of either childcare benefits receivers or students. However, this may have negative consequences for the prediction accuracy and may breed a new type of bias towards certain groups and individuals within the minority groups. As such, there is a way, but it would imply more experimentation with algorithms and protected characteristics, which may not be advisable in practice considering the potential consequences.

These impossible scenarios of algorithmic governance reflect the “analog” observation that the blindness approach to equality of treatment is limited, because true blindness is unachievable, and protected characteristics are indistinguishably interconnected with other individual characteristics. On the other hand, the awareness approach to equality of outcomes is also limited, because it can have unpredictable consequences on the accuracy of predictions. The choice between programming blindness or awareness into algorithms reflects the long-standing political dilemma between equality and equity, which traditionally divides experts. Once again, the technological quick fix falls short of solutionist promises, as this dilemma is notoriously difficult and contextual. The dilemma still exists, even if hidden behind the opaque self-learning mechanisms, and algorithmizing governance only makes public discourse around these issues more inaccessible. This is often the case when technological solutions are offered to problems that are inherently political, economic, and ethical.³⁷ Therefore, one cannot be cautious enough when trying to algorithmize governance, particularly with regard to characteristics related to historical systemic discrimination, e.g. on the basis of race and ethnicity. In other words, achieving algorithmic fairness in a setting that has long roots in structural discrimination may not be possible,³⁸ which is a clear argument against the experimentation with algorithmic governance on vulnerable members of society, as this perpetuates the cycle of discrimination and vulnerability.

4 Trust

For a concept so ubiquitous that it seems almost self-explanatory, trust is defined in a surprising variety of ways: as a propensity or a personality characteristic in psychology,³⁹ a relational characteristic or a characteristic of a social network in sociology,⁴⁰ a valuative orientation towards the polity in political science,⁴¹ and as

³⁷ See Sadowski and Selinger (2014).

³⁸ See Russell (2021).

³⁹ See Glanville and Paxton (2007).

⁴⁰ See Cook and Santana (2020).

⁴¹ See Levi and Stoker (2000).

the key ingredient of economic development in economics.⁴² As such, trust is a multidimensional construct, likely determined simultaneously by psychological and structural dispositions, relational characteristics and the greater societal, economic and cultural context. Since all of the above-listed determinants withstood empirical scrutiny and the test of time, it follows that trust is at the same time dispositional, relational, and situational. In this chapter, the focus is on the relational dimension of trust, mainly in the relation government—citizens, that comes to the forefront in considerations of algorithmic governance.

In this chapter, two types of trust are of interest with regard to algorithmic governance, namely political (trust in different sections of the government) and social trust (in other people). Although political and social trust are distinct constructs, they are undoubtedly related, and their relationship is conceptualized in a variety of theoretical mechanisms depending on the discipline.⁴³ Empirically, however, the concepts are strongly associated, which lead Zmerli and Newton⁴⁴ to conclude that social and political trust simultaneously overlap and reinforce each other. Indeed, longitudinal research provides indications that political trust exerts positive effects on social trust,⁴⁵ and also that social trust exerts positive effects on political trust.⁴⁶ Therefore, the empirical findings suggest that political trust is best defined as partly a subset of social trust and partly independent of it, and the two types of trust can affect one another through a continuous feedback loop between the two.

Political trust is commonly acknowledged as essential for democracies to function effectively.⁴⁷ However, a decline in political trust has been observed consistently and chronically across the board in recent decades.⁴⁸ A wide array of possible explanations is offered for the observed decline, which makes it difficult to recommend policies that would address it. Namely, political science literature identifies endogenous and exogenous causes of declining political trust. Endogenous explanations attribute the decline in political trust to the performance of public institutions, for example political scandals, or more generally functioning of the government and policies; while exogenous attribute the decline to factors other than the performance of public institutions, such as overarching cultural or economic trends.⁴⁹ In the Netherlands specifically, Bovens and Wille⁵⁰ identify five plausible exogenous causes of the observed decline in political trust, namely the deteriorating economy, changing role of the media, developments in citizens' values and expectations, generational shifts, and the loss of social capital. As these and similar explanations are

⁴² See Cohen (2020).

⁴³ See Newton et al. (2018).

⁴⁴ See Zmerli and Newton (2008).

⁴⁵ See Tao et al. (2014).

⁴⁶ See Newton and Zmerli (2011); Bargsted et al. (2022).

⁴⁷ See Catterberg and Moreno (2006); Keymolen and Voorwinden (2020).

⁴⁸ See Dogan (2005); Dassonneville and McAllister (2021).

⁴⁹ See Mishler and Rose (2001).

⁵⁰ See Bovens and Wille (2008).

to a large extent independent of the functioning of the government itself, some authors argue that political trust is not an indicator of a good democracy, and its decline should not be of concern for governments.⁵¹ However, since both the endogenous and the exogenous lines of reasoning received empirical support, it follows that governments would be best advised to focus on building trust through the endogenous mechanisms at their disposal, such as preventing political scandals and focusing on improving the perception of performance through transparency, which may help mitigating the adverse effects of exogenous factors on political trust. Moreover, even exogenous causes, such as global health or financial crises, can be endogenously handled in various ways, some more conducive to political trust than others. Algorithmic governance, due to its experimental character, already at first glance does not appear conducive to building political trust, but let us return to the Dutch fraud detection examples to illustrate why.

In cases when governments use algorithmic surveillance, e.g. for fraud detection, the relational dimension (i.e. reciprocity) of trust is highlighted. Namely, the mere use of algorithmic surveillance for the purpose of fraud detection signals distrust that the government harbors towards the vulnerable citizens who received benefits. Learning that government distrust them enough to experiment with algorithmic surveillance to prevent fraud may in turn decrease vulnerable citizens' political trust. At the same time, for reasons likely including precisely the relational nature of trust, political trust is commonly found to be lower among individuals of lower socio-economic status,⁵² i.e. socially vulnerable citizens. Acknowledging the relational nature of trust makes apparent the vicious cycle of how failures of algorithmic governance, such as biases that lead to discrimination, can exacerbate and perpetuate the cycle of mistrust, further deteriorating already low trust among socially vulnerable populations. Additionally, when policy experimentation with algorithmic governance backfires, such as when entire governments resign over the issue, the trust is not only damaged between the government and those who were "test subjects" of policy experimentation, but also between the government and the rest of the population observing the outcomes of algorithmic governance experimentation.

Poverty (and, by extension, social vulnerability) is traditionally associated with lower levels of political trust.⁵³ At the same time, concerns have been raised that algorithmic governance may further erode trust.⁵⁴ It follows, thus, that experimentation with algorithmic governance targeting vulnerable communities ought to be regarded with caution. However, there have been recent examples where caution was lacking, and the consequences for affected vulnerable communities have been grave. Apart from the Dutch "xenophobic algorithms" for fraud detection, there are also reports on the Michigan Integrated Data Automated System for detection of

⁵¹ See Stimson (2004).

⁵² See Goubin and Hooghe (2020).

⁵³ See Papadakis et al. (2017); MacNabb (2022).

⁵⁴ See Meijer and Grimmlikhuijsen (2020).

unemployment fraud, which had a ninety-three percent error rate.⁵⁵ Unfortunately, concerning examples of algorithmic governance rollout targeting vulnerable communities are not limited to fraud detection. Other examples include the case of the Los Angeles Police Department's increasing surveillance in certain neighborhoods indicated to have high risk of crime by predictive Big Data analytics;⁵⁶ and the use of Allegheny Family Screening Tool in Pennsylvania, which resulted in compounding the effects of systemic racial discrimination;⁵⁷ the use of biometric facial recognition technology (FRT) as a form of algorithmic surveillance used for policing protesters in 2020, during the Black Lives Matter protests;⁵⁸ or current controversial algorithmic governance practices relating to border control, which commonly include predictive analytic algorithms for risk calculation, dialect recognition for asylum applications, mobile data analysis, and algorithmic surveillance through GPS tagging and facial recognition.⁵⁹ Many of these examples lack the reversibility and evidence based qualifiers for policy experimentation proposed by Listokin.⁶⁰ Although there is no data available to confirm, it would hardly be surprising if in these cases the rollout of algorithmic governance resulted in further erosion of trust among the affected vulnerable communities. Moreover, these examples substantiate Lyon's⁶¹ observation that some individuals, groups and areas are more often subjected to (algorithmic) surveillance than others, revealing as such the connectedness of the values of equality and trust, whereby a lack of trust from the government's side reinforces inequality, and increased inequality further deteriorates trust on the citizens' side.

In sum, it is likely that both endogenous and exogenous factors affect political trust, which then affects social trust. At the same time, political trust affects a myriad of other outcomes, including the ones concerning the government, but also the ones concerning the citizens, both with regards to outcomes and subjective experiences. Therefore, it is advisable for governments to invest efforts in increasing trust through optimizing the endogenous factors, which would then have a ripple effect on the interconnected network of indicators of a healthy society, including political and social trust.

⁵⁵ See Rankin (2023).

⁵⁶ See Brayne (2017).

⁵⁷ See Gerchick et al. (2023).

⁵⁸ See Powell (2022).

⁵⁹ See Ozkul (2023).

⁶⁰ See Listokin (2008).

⁶¹ See Lyon (2003).

5 Discussion and Conclusion

The central question of this chapter relates to the societal antecedents and consequences of novel algorithmic governance policies commonly tested on vulnerable populations. With regards to antecedents, two complementary conclusions stand out. Firstly, in the worst-case scenario, it appears that algorithmic governance is at the same time a reflection of, and an active agent in perpetuating and creating power disparities, as it is often deployed to surveil and control vulnerable populations. Moreover, the policies are implemented with the goal of testing and developing the technology itself, whereby vulnerable groups and individuals are used as convenient experimentation subjects with regard to traditional power asymmetries. Secondly, in the best-case scenario, algorithmic governance is an example of technological solutionism, wherein technology is regarded as the only viable solution, and as such technology guides the development of policy, rather than vice-versa.⁶² Either way, the examples of failures of algorithmic surveillance that were uncovered only after leaving a lasting mark on the lives of socially vulnerable individuals they were tested on are far too common. This chapter reflects on notable examples of failures of algorithmic governance experimentation targeting vulnerable populations in the domains of fraud detection, border control, and the criminal judicial system. These failures reflect the limitations of technological solutions, as they do nothing to address the systemic reasons behind social vulnerability and marginality that motivate governments to control these communities through means of algorithmic surveillance in the first place.⁶³ Moreover, the biased predictions of an algorithm may become self-fulfilling, creating thereby a vicious cycle and perpetuating historical biases.⁶⁴

Concerning public values, this chapter has shown how the premature and irresponsible rollout of algorithmic governance targeting vulnerable communities can have detrimental effects on public values. In particular, this chapter addresses the effects of algorithmic governance failures on the public values of equality and trust. However, a myriad of additional public values is potentially affected, such as transparency, accountability, compassion, and freedom. Future research could therefore examine the effects on these and other potentially affected public values, with an emphasis on the interconnectedness of values. This notion of interconnectedness is also evident from the vicarious learning consideration that emerged from the conceptual analysis of the presented examples. Not only do the unpredictable effects and mistakes brought on by the experimental implementation of algorithmic governance deprive the identified public values among the targeted vulnerable groups, but they also, through vicarious learning, deprive these values among the general population of citizens who were not directly affected. Although experts warn about these and other issues, algorithmic governance is already well underway, and is likely

⁶² See Morozov (2013).

⁶³ See Herbert et al. (2018); Molnar (2021).

⁶⁴ See Sahlgren (2020).

here to stay despite the fiascos described in this chapter. What is left to do is “salvaging whatever we can” through advocating for, as much as possible, ethical experimentation with emerging technologies in governance, notably by accounting for social vulnerability as a counterindication for sociotechnical experimentation.

Furthermore, in light of this analysis, it is imperative to maintain a conservative approach to the scale of sociotechnical experimentation. Acknowledging the possibility of unintended adverse effects of, for example, algorithmic governance experimentation necessitates a piecemeal approach whereby small-scale, gradual experimentation is conducted, emphasis is placed on learning from these tests, and the scale of experimentation is incrementally increased depending on the analysis of earlier small scale implementation. This approach emphasizes the precautionary principle (minimizing harm), and is as such hardly a revolutionary idea, since it has been recognized as a good practice in policy implementation more than 60 years ago,⁶⁵ and even then it echoed ideas on incremental policy change outlined by Burke as far back as eighteenth century.⁶⁶ While holistic and abrupt changes in policy can in some cases be desirable and even necessary, I would argue that top-down implementation of emerging technologies by governments to cut costs and/or increase control is decisively not one of those cases. This argument is substantiated by the oftentimes irreversible nature of policies relying on emerging technologies, despite the irreversibility of policies having been identified as a counterindication for policy experimentation.⁶⁷ As can be seen from examples throughout this chapter, while the algorithmic governance policies themselves can be revoked, the consequences such rushed policies can have on individuals, and in particular on vulnerable individuals, cannot.

In conclusion, social vulnerability should be a clear counterindication for algorithmic governance experimentation, particularly with regard to systems of surveillance and profiling. On the contrary, what characterizes suitable candidates for testing governance by emerging technologies is precisely the opposite—namely, resilience, defined as the ability to successfully adapt and overcome adversities.⁶⁸ One may wonder, would it not be wiser to test the new modes of algorithmic governance on legal subjects who have the resilience, i.e. the power and resources, to soften the blow of potential mistakes and injustices that constitute inherent risks in experimental governance systems? In the midst of the Dutch childcare benefits scandal, for example, thousands of children were taken into foster care, and some citizens tragically committed suicide, due to the economic hardships that wrongly issued fines put the affected families through. The argument could be made that, should algorithms make mistakes in detecting, for example, fraud and crime performed by corporations, rather than natural persons, these mistakes would be less consequential. Moreover, private companies, in particular multinationals who are

⁶⁵ See Popper (1961).

⁶⁶ See Burke (2016).

⁶⁷ See Listokin (2008).

⁶⁸ See Masten (2001).

often implicated in white-collar crime allegations, typically have the means to fight legal battles in pursuit of justice, which is often not the case for the socially vulnerable. In other words, if we must find out that algorithms for crime detection are faulty, it is arguably safer to find that out through experimentation targeting for-profit companies, than natural persons who are, more often than not, already in a socially vulnerable position. An illustrative example of successful use of algorithms in detection of white-collar crime is the use of “AI lawyer”, a self-learning algorithm for reviewing documents in the lawsuit against Rolls Royce regarding bribery and corruption.^{69,70} The process resulted in Rolls Royce paying out a record settlement of £671 million.

The Rolls Royce example, while perhaps inviting to question the usefulness of settlements as a tool for disciplining companies that can easily afford such payouts, also clearly illustrates the ability of corporations to resiliently deal with penalties. Conversely, when algorithmic governance is used to surveil and control vulnerable populations, it dehumanizes these individuals and hinders the standards for due process and justice.⁷¹ As demonstrated in this chapter, experimentation with the use of emerging technologies in governance of vulnerable individuals and communities can have grave consequences of existential nature. It is therefore far more appropriate to experiment with the use of emerging technologies in governance on legal entities that enjoy more socio-economic power and, consequently, more resilience, to deal with the potentially inaccurate outcomes through, for example, pursuing justice through the legal system. Moreover, the potential negative effects are disproportionate for vulnerable communities with regards to public values, as the conducted value analysis showed social vulnerability is already inversely associated with the values of equality and trust. As such, experimentation with algorithmic governance likely further undermines these already depleted values, which may cause a compounding effect and a plummeting sense of equality and trust among vulnerable communities. Therefore, careful and context-specific value analysis should be the basis of any algorithmic governance policy, and social vulnerability should be considered a counterindication for experimentation.

Acknowledgements I would like to thank Esther Keymolen, Sofia Ranchordás and Malou Beck for their valuable feedback throughout the writing process, as well as the participants of the “Public Values and Emerging Technologies – Values, Trust and Compliance by Design” workshop in Utrecht, 2024.

⁶⁹ See Serious Fraud Office (2017).

⁷⁰ Here it is worth noting another key distinction between the use of self-learning algorithms for the Rolls Royce case and for the childcare benefits fraud case, besides the target. Namely, in the former case, the algorithm was used reactively, to analyze data of an already identified target under investigation, whereas in the latter case it was used proactively, in the form of predictive analysis to identify targets.

⁷¹ See Valentine (2019).

References

- Andrejevic M (2014) Big data, big questions: the big data divide. *Int J Commun* 8:1673–1689. Available at: <https://ijoc.org/index.php/ijoc/article/view/2161>
- Angwin J, Larson J, Mattu S, Kirchner L (2016, May 23) Machine bias: there's software used across the country to predict future criminals. And it's biased against blacks. ProPublica. <https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>
- Ansell, C. K., & Bartenberger, M. (2016). Varieties of experimentalism. *Ecological Economics*, 130, 64–73.
- Appleman LI (2020) The captive lab rat: human medical experimentation in the carceral state. *Boston Coll Law Rev* 61(1):1–69. <https://heinonline.org/HOL/P?h=hein.journals/bclrl61&i=7>
- Bargsted M, Ortiz C, Cáceres I, Somma NM (2022) Social and political trust in a low trust society. *Polit Behav* 1:1401–1420. <https://doi.org/10.1007/s11109-021-09762-2>
- Bikundo E (2020) Experimental norms: power-knowledge, bare life and medical trials. *Law Technol Hum* 2(2):42–52. <https://search.informit.org/doi/10.3316/informit.563367904618129>
- Bovens M, Wille A (2008) Deciphering the Dutch drop: ten explanations for decreasing political trust in the Netherlands. *Int Rev Adm Sci* 74(2):283–305. <https://doi.org/10.1177/0020852308091135>
- Brayne S (2017) Big data surveillance: the case of policing. *Am Sociol Rev* 82(5):977–1008. <https://doi.org/10.1177/0003122417725865>
- Burke E (2016) Reflections on the revolution in France. In: Blaug R, Schwarzmantel J (eds) *Democracy: a reader*. Columbia University Press, pp 234–238
- Catterberg G, Moreno A (2006) The individual bases of political trust: trends in new and established democracies. *Int J Public Opin Res* 18(1):31–48. <https://doi.org/10.1093/ijpor/edh081>
- Cohen MA (2020) Trust in economy. In: Simon J (ed) *The Routledge handbook of trust and philosophy*. Routledge, pp 283–297
- Collingridge, David. 1992. *The Management of Scale: Big Organizations, Big Technologies, Big Mistakes*. London: Unwin Hyman.
- Commission. https://commission.europa.eu/strategy-and-policy/priorities-2019-2024/europe-fit-digital-age/europes-digital-decade-digital-targets-2030_en
- Cook KS, Santana JJ (2020) Trust: perspectives in sociology. In: Simon J (ed) *The Routledge handbook of trust and philosophy*. Routledge, pp 189–204
- Cullen BT, Pretes M (2000) The meaning of marginality: interpretations and perceptions in social science. *Soc Sci J* 37(2):215–229. [https://doi.org/10.1016/S0362-3319\(00\)00056-2](https://doi.org/10.1016/S0362-3319(00)00056-2)
- Dassonneville R, McAllister I (2021) Explaining the decline of political trust in Australia. *Aust J Polit Sci* 56(3):280–297. <https://doi.org/10.1080/10361146.2021.1960272>
- de Heer, S. (2023). A Scandal on AI in Administration, Again. *Verfassungsblog*.
- Dewey J (1938) *Logic, the theory of inquiry*. Holt
- Dogan M (2005) Erosion of confidence in thirty European democracies. *Comp Sociol* 4(1–2):11–54. <https://doi.org/10.1163/1569133054621932>
- Ersoy S, van der Gaag S (2023, June 21) Studenten met migratieachtergrond opvallend vaak beschuldigd van fraude, minister wil systeem grondig nagaan. NOS. <https://nos.nl/op3/artikel/2479700-studenten-met-migratieachtergrond-opvallend-vaak-beschuldigd-van-fraude-minister-wil-systeem-grondig-nagaan>
- Eubanks, V. (2018). *Automating inequality: How high-tech tools profile, police, and punish the poor*. St. Martin's Press.
- EuropeanCommission(2024)Europe'sDigitalDecade:digitaltargetsfor2030.EuropeanCommission. https://commission.europa.eu/strategy-and-policy/priorities-2019-2024/europe-fit-digital-age/europes-digital-decade-digital-targets-2030_en
- Fineman, M. A. (2008). The vulnerable subject: Anchoring equality in the human condition. In *Transcending the boundaries of law* (pp. 177–191). Routledge-Cavendish.
- Fordham M, Lovekamp WE, Thomas DS, Phillips BD (2013) Understanding social vulnerability. In: Thomas DSK, Brenda D (eds) *Social vulnerability to disasters*, 2. CRC Press, pp 1–29

- Gerchick M, Jegede T, Shah T, Gutierrez A, Beiers S, Shemtov N, Xu K, Samant A, Horowitz A (2023) The devil is in the details: interrogating values embedded in the Allegheny Family Screening Tool. 2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT '23), pp 1292–1310
- Glanville JL, Paxton P (2007) How do we learn to trust? A confirmatory tetrad analysis of the sources of generalized trust. *Soc Psychol Q* 70(3):230–242. <https://doi.org/10.1177/019027250707000303>
- Goubin S, Hooghe M (2020) The effect of inequality on the relation between socioeconomic stratification and political trust in Europe. *Soc Justice Res* 33:219–247
- Heikkilä M (2022, March 29) Dutch scandal serves as a warning for Europe over risks of using algorithms. *Politico*. <https://www.politico.eu/cdn.ampproject.org/c/s/www.politico.eu/article/dutch-scandal-serves-as-a-warning-for-europe-over-risks-of-using-algorithms/amp>
- Herbert S, Beckett K, Stuart F (2018) Policing social marginality: contrasting approaches. *Law Soc Inq* 43(4):1491–1513. <https://doi.org/10.1111/lsi.12287>
- Hoger Onderwijs Persbureau (2024, March 27) House of representatives apologies after duo scandal how about compensation. <https://www.erasmusmagazine.nl/en/2024/03/27/house-of-representatives-apologies-after-duo-scandal-how-about-compensation/>
- Huitema D, Jordan A, Munaretto S, Hildén M (2018) Policy experimentation: core concepts, political dynamics, governance and impacts. *Policy Sci* 51:143–159. <https://doi.org/10.1007/s11077-018-9321-9>
- Keymolen E, Voorwinden A (2020) Can we negotiate? Trust and the rule of law in the smart city paradigm. *Int Rev Law Comput Technol* 34(3):233–253. <https://doi.org/10.1080/13600869.2019.1588844>
- Kleinberg J, Mullainathan S (2019) Simplicity creates inequity: implications for fairness, stereotypes, and interpretability. *Proceedings of the 2019 ACM Conference on Economics and Computation*, pp 807–808
- Krohn W, Weyer J (1994) Society as a laboratory: the social risks of experimental research. *Sci Public Policy* 21(3):173–183. <https://doi.org/10.1093/spp/21.3.173>
- Kopp, O. (2006). Historical review of unethical experimentation in humans. *Ethics Professions*, 2.
- Levi M, Stoker L (2000) Political trust and trustworthiness. *Annu Rev Polit Sci* 3(1):475–507. <https://doi.org/10.1146/annurev.polisci.3.1.475>
- Lipton Z, McAuley J, Chouldechova A (2018) Does mitigating ML's impact disparity require treatment disparity?, arXiv:1711.07076, pp 1–11
- Listokin Y (2008) Learning through policy variation. *Yale Law J* 118(3):480–553. https://heinonline.org/HOL/Page?handle=hein.journals/ylr118&div=18&g_sent=1&casa_token=ptgzd4pYAYUAAAAA:0vvULuiwEp1cpF6pqkNT2GUfLL17vYoRMiq9X7D4x-ADGMst2SYs_V2u_zZcxmizUW0SR7CQ&collection=journals
- Luna, F. (2009). Elucidating the concept of vulnerability: Layers not labels. *IJFAB: International Journal of Feminist Approaches to Bioethics*, 2(1), 121–139.
- Lyon D (2003) *Surveillance as social sorting: privacy, risk, and digital discrimination*. Psychology Press
- Mackenzie, C. (2014). The importance of relational autonomy and capabilities for an ethics of vulnerability. *Vulnerability: New essays in ethics and feminist philosophy*, 33–59.
- MacNabb E (2022) Social vulnerability and trust in government disaster response: the case of a potential flood event in a metropolitan area. University of Nebraska at Omaha
- Madden M, Gilman M, Levy K, Marwick A (2017) Privacy, poverty, and big data: a matrix of vulnerabilities for poor Americans. *Wash Univ Law Rev* 95(1):53–125. https://heinonline.org/HOL/Page?handle=hein.journals/walq95&div=5&g_sent=1&casa_token=T6JUm10qU9cAAAA:wPcF1ZowW8L-1yowIbPnqw299MhJ64o2atpzk6to0aVAqftXKPG2_1Cah8rUHidrvWmqJPQdog&collection=journals
- Masten AS (2001) Ordinary magic: resilience processes in development. *Am Psychol* 56(3):227–238
- McFadgen B, Huitema D (2018) Experimentation at the interface of science and policy: a multi-case analysis of how policy experiments influence political decision-makers. *Policy Sci* 51:161–187. <https://doi.org/10.1007/s11077-017-9276-2>

- Meijer A, Grimmelikhuijsen S (2020) Responsible and accountable algorithmization: how to generate citizen trust in governmental usage of algorithms. In: Schuilenburg M, Peeters R (eds) *The algorithmic society: technology, power, and knowledge*. Routledge, pp 53–66
- Metcalfe J, Crawford K (2016) Where are human subjects in big data research? The emerging ethics divide. *Big Data Soc* 3(1):1–14. <https://doi.org/10.1177/2053951716650211>
- Minow M (2023) Equality, equity, and algorithms: learning from Justice Rosalie Abella. *Univ Toronto Law J* 73(Supplement 2):163–178. <https://doi.org/10.3138/utlj-2023-006>
- Mishler W, Rose R (2001) What are the origins of political trust? Testing institutional and cultural theories in post-communist societies. *Comp Polit Stud* 34(1):30–62. <https://doi.org/10.1177/0010414001034001002>
- Molnar P (2021) Technological testing grounds and surveillance sandboxes: migration and border technology at the frontiers. *Fletcher Forum World Aff* 45(2):109–118. https://heinonline.org/hol-cgi-bin/get_pdf.cgi?handle=hein.journals/forwa45&section=24&casa_token=pcMHmIyFInkAAAAA:mbFwMdXhw6pYN2ct8W_smec1dEVYwpNI7AAkMYWw4oQ-o9NWz3tpyNXCQghH2sfnpw--oQmFbQ
- Morozov E (2013) To Save Everything, Click Here: The Folly of Technological Solutionism. PublicAffairs
- Negri S (2017) Unethical human experimentation in developing countries and international criminal law: old wine in new bottles? *Int Crim Law Rev* 17(6):1022–1048. <https://doi.org/10.1163/15718123-01761352>
- Newton K, Zmerli S (2011) Three forms of trust and their association. *Eur Polit Sci Rev* 3(2):169–200. <https://doi.org/10.1017/S1755773910000330>
- Newton K, Stolle D, Zmerli S (2018) Social and political trust. In: Uslander EM (ed) *The Oxford handbook of social and political trust*. Oxford University Press
- Ozkul D (2023) Automating immigration and asylum: the uses of new technologies in migration and asylum governance in Europe. Oxford Department of International Development
- Papadakis N, Papargyris A, Dafermos V, Basta M, Kyridis A, Drakaki M, Plymakis S (2017) Youth and NEETs in Greece, within the crisis era. Social vulnerability, unemployment, public trust and issues of political behavior. *Youth in Crisis Countries-Life Situations and Political Attitudes of Adolescents in Southern/Eastern Europe and North Africa* 12:114–130
- Parlementaire ondervragingscommissie Kinderopvangtoeslag (2020). Ongekend onrecht. (Kamerstukken II 2020/21, 35510, nr. 2)
- Pelc S (2017) Marginality and marginalization. In: Chand R, Nel E, Pelc S (eds) *Societies, social inequalities and marginalization: marginal regions in the 21st century*. International Mountain Society, pp 13–28
- Popper KR (1961) *The poverty of historicism*. Routledge & Kegan Paul
- Powell LC (2022) The good, the bad, and the ugly: Black Lives Matter Protests, the January 6th Insurrection, and Facial Recognition Technology as Admissible Evidence. *Am Univ Law Rev* 72(1):277–335. https://heinonline.org/HOL/Page?handle=hein.journals/aulr72&div=10&g_sent=1&casa_token=fEpnVmuiQe8AAAAA:KnP4JOqEWpMeX2lRW_oITlUvi_pO1u16GJJ6TmMhtnnzj_zAiJ4XQL0lzUNTlcz-oNAjDR4Dg&collection=journals
- RankinSMG(2023)TheMiDASTouch:Atuahene’s“Stategraft”andUnregulatedArtificialIntelligence. N Y Univ Law Rev Online 98:225–245. https://heinonline.org/HOL/Page?handle=hein.journals/nyulro98&div=11&g_sent=1&casa_token=8qrogZ-JKnUAAAAA:5mtcrjmq6DVRB-lvNoig823DEAMXzU8pEvOL2HQOr7Cii-Ft6n1Uv66wrE2ItsR6-tA2N2iFZA&collection=journals
- Rodriguez M (2022) Equity vs. equality in automated decision-making systems. New York University
- Russell J (2021) Machine learning equity and accuracy in an applied justice setting. 2021 IEEE International Conference on Smart Computing (SMARTCOMP), pp 215–221
- Sadowski J, Selinger E (2014) Creating a taxonomic tool for technocracy and applying it to Silicon Valley. *Technol Soc* 38:161–168
- Sahlgren, O. (2020). *Algorithmic Decision-making, Discrimination and Disrespect: An Ethical Inquiry* (Master’s thesis).

- Sandvik KB, Jacobsen KL, McDonald SM (2017) Do no harm: a taxonomy of the challenges of humanitarian experimentation. *Int Rev Red Cross* 99(904):319–344. <https://doi.org/10.1017/S181638311700042X>
- Serious Fraud Office (17 January 2017) SFO completes £497.25m Deferred Prosecution Agreement with Rolls-Royce PLC. Serious Fraud Office. <https://www.sfo.gov.uk/2017/01/17/sfo-completes-497-25m-deferred-prosecution-agreement-rolls-royce-plc/>
- Simons J, Adams Bhatti S, Weller A (2021) Machine learning and the meaning of equal treatment. *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pp 956–966
- Spielman SE, Tuccillo J, Folch DC, Schweikert A, Davies R, Wood N, Tate E (2020) Evaluating social vulnerability indicators: criteria and their application to the Social Vulnerability Index. *Nat Hazards* 100:417–436. <https://doi.org/10.1007/s11069-019-03820-z>
- Stimson JA (2004) *Tides of consent: how public opinion shapes American politics*. Cambridge University Press
- Tao R, Yang DL, Li M, Lu X (2014) How does political trust affect social trust? An analysis of survey data from rural China using an instrumental variables approach. *Int Polit Sci Rev* 35(2):237–253. <https://doi.org/10.1177/0192512113492599>
- Tassey G (2014) Innovation in innovation policy management: the Experimental Technology Incentives Program and the policy experiment. *Sci Public Policy* 41(4):419–424. <https://doi.org/10.1093/scipol/sct060>
- Taylor, L. (2021). Exploitation as innovation: research ethics and the governance of experimentation in the urban living lab. *Regional Studies*, 55(12), 1902–1912.
- ten Seldam B, Brenninkmeijer A (2021, April 30) The Dutch benefits scandal: a cautionary tale for algorithmic enforcement. *EU Law Enforcement*. <https://eulawenforcement.com/?p=7941>
- Valentine S (2019) Impoverished algorithms: misguided governments, flawed technologies, and social control. *Fordham Urb Law J* 46(2):364–427. https://heinonline.org/hol/cgi-bin/get_pdf.cgi?handle=hein.journals/frdurb46&ion=14&casa_token=p_DFmbIKv_sAAAAA:zUMDY0XOIPKKFNWnIFBCeqM5DKcWBJkRiG9OndEb1xHqSWbLxuf-76VskHpsqnUkPTUG4Fu8200
- van de Poel I (2016) An ethical framework for evaluating experimental technology. *Sci Eng Ethics* 22(3):667–686. <https://doi.org/10.1007/s11948-015-9724-3>
- van Zoonen L (2020) Data governance and citizen participation in the digital welfare state. *Data Policy* 2:e10-1–e10-17. <https://doi.org/10.1017/dap.2020.10>
- Zmerli S, Newton K (2008) Social trust and attitudes toward democracy. *Public Opin Q* 72(4):706–724. <https://doi.org/10.1093/poq/nfn054>

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.



The Frameworks of Trust and Trustlessness Around Algorithmic Control Technologies: A Lost Sense of Community



Balázs Bodó and Linda Weigl

Abstract Certain techno-political infrastructures, e.g. blockchains, aim to replace our existing social and institutional modes of producing trust as a social resource. Can they successfully do that, without the reliance on the very same institutions, which could safeguard and guarantee their trustworthiness in the first place? By now we have more than a decade of experience trying to build autonomous, code-driven, private ordering infrastructures, designed to complement, disrupt, or replace both private and public institutions. The revolution of these ‘trustless’ digital technologies is yet to happen, raising concerns about their promises to address the existing trust challenges of centralized institutions, their capacity to eliminate the societal reliance on trust, and the potential consequences thereof. Therefore, in this chapter, we pose the following questions: How does trustlessness through the elimination of more-or-less trusted middlemen impact our values and our sense of belonging? How does the decision to end trust maintenance through trustless technologies impact the cultivation of a sense of community within a society? This chapter addresses these questions by critically reviewing the claims surrounding the trustlessness of automated, code-as-law-based governance systems in the field of digital identity management—an area that continues to command the attention of various organizations and institutions.

1 Algorithmically Produced Trust and Trustlessness

When one party places trust in another, they rely on the trusted party’s intentions and capabilities to act in their best interest, and thereby expose themselves to the possibility of harm, deception, or failure. This inherent vulnerability is what makes trust a social construct and human phenomenon. Algorithmic control systems, such as blockchain technologies are developed to automate tasks and thereby reduce human intervention (and errors) to a minimum. They optimize resources by

B. Bodó (✉) · L. Weigl

Institute for Information Law, University of Amsterdam, Amsterdam, Netherlands

e-mail: b.bodo@uva.nl; l.weigl@uva.nl

© The Author(s) 2025

J. Goossens et al. (eds.), *Public Governance and Emerging Technologies*,
https://doi.org/10.1007/978-3-031-84748-6_16

355

reducing invested time, money, and effort, increase efficiency, and distribute decision-making and power away from traditional institutional entities. While some view them as progressive alternatives to the perceived failures and untrustworthiness of traditional legal systems and institutions, others perceive them as strategic means to disrupt the existing institutional order to further their narrow, self-interested economic, political, or ideological agenda. What unites both contrasting perspectives, regardless of their ability to automatize, economize or decentralize activities, is their evaluation of the technology's capacity to reshuffle power dynamics in society by reconfiguring, remediating, and potentially replacing institutional and interpersonal trust relations.

This replacement of trust relations is achieved by two means. First, a trustless system is able to generate an immutable, consensus-driven, and publicly accessible ledger of transactions without the involvement of a centralized intermediary.¹ Second, its 'trustlessness' is embodied by rules written as self-enforcing code running on the blockchain base layer, enforcing the compliance of anonymous parties through automated, algorithmic means.

Both these specificities of blockchain-based systems stem from a particular vision on trust in individuals and institutions. On the one hand, individuals' trustworthiness is inscrutable because their identity is unknown, and no identification is needed to take part in blockchain-based transactions. If, however, the trustworthiness of the individual behind the transaction is not possible to establish, the system has to offer other safeguards and guarantees to ensure that transactions take place even though the counterparty is anonymous—hence the need for strong, self-enforcing code-based rules. Confidence (i.e.: the expectation that the system will work as expected) in algorithmic systems is derived from the predictability of the mathematical rigor of the hashing algorithm, in particular public-private key cryptography, as well as the economic incentive schemes for miners and consensus algorithms governing the network. Hence, when users decide to subject themselves to the authority of a technological system, they do so because they may have confidence in its predictable, mechanistic, 'objective' operation, where no trust is needed, because no possible (human or institutional) betrayal is possible. These confidence-enhancing features, techno-optimists argue, allow the technology to enhance decentralized decision-making, transparency, tamper-resistance, automation, impartiality, and objectivity. The goal is to eliminate human malintent and fallibility, corruptible institutions, arbitrary enforcement, and essentially remove human and institutional messiness from the system altogether.²

¹We acknowledge that all decentralized systems rely to some extent on centralized components to function, such as miners, validators, developers, computing infrastructure providers, developers, etc. The trustworthiness of these actors is of course an important question, but not substantially different from the trustworthiness concerns of other centralized institutions.

²However, despite the high degree of confidence that algorithmic control technologies can qualify for, it can be argued that the requirement for trust *in* such systems persists, thereby challenging the system's allegedly trustless character. This is due to their socio-technical nature, as technology can never be entirely impartial, unbiased, or apolitical. For a more detailed elaboration see Bijker et al.

Public and permissionless blockchains, known as such ‘trust-free’ or ‘trustless’ systems, pledge to transform interactions among anonymous-by-design peers that typically rely on trust, often mediated by third-party providers.³ As we discuss in more detail later, identity information, and the way it is produced, maintained, made available, or not, is a central question that defines interpersonal and impersonal trust relations, as well as the socio-technical and institutional modes of trust production. Second, the deep distrust towards traditional, rule-based, socially embedded intermediaries requires that some of the functions they originally fulfill—such as maintaining ledgers of all sorts—be provided through other approaches, such as reliance on algorithmically defined rules. This raises the question of *what happens to our social trust relations organized around the principles of anonymity and algorithmic control?*

The claim of algorithmic objectivity is not limited to the domain of blockchain-based technologies. Other automated systems, such as AI, and algorithmic decision-making systems also operate on the principle of quantified objectification: on the claim that once unbiased, comprehensive data is fed into objectively formulated, automatically followed transparent algorithmic rules-based systems, the resulting output would therefore be more trustworthy than those produced by institutional processes riddled with human error, judgement, and subjectivity.⁴ If the algorithmic decision-making system uses people as its input, then individuals (their identity, history, values, decisions, social networks) also get transformed by measurement, and conversion into data points, and the complexity of their trustworthiness gets simplified into plain scores of creditworthiness, fraud risk indicators, or online reputation.

Blockchains, AI, and algorithmically governed platforms are just a few examples of techno-social systems that aim to produce trust in various social, economic, cultural, or interpersonal relations.⁵ Here is why: irrespective of the actual trustworthiness or reliability of algorithmic trust-producing systems, their mere existence has an immediate impact: for better or worse they disrupt the trust in external actors, individuals, and traditional institutions. Blockchain-based systems, for example, don’t simply try to eliminate the need for trust, but *try to eliminate the social, institutional practices of traditional trust production*. While not inherently trustless, those technologies can be viewed as actively contributing to the cultivation of societal and interpersonal trustlessness, both through removing the need for trust, and through the elimination of centralized intermediaries and institutions, together with their trust-producing functions. The question, therefore, is not just about how human actors engaging in trust-necessitating transactions within the system trust each other under the new, technologically defined conditions. It is about how trust between

(1989) and Hughes (1987). Confidence in blockchain technologies, for instance, may thus be challenged by the decisions of human actors in the system, such as influential core developers working on the blockchain protocol, or miners. See De Filippi et al. (2020).

³ See Beck et al. (2016) and Hawlitschek et al. (2018).

⁴ See Christin (2016).

⁵ See Bodó (2021a, b), Botsman (2017), and Keymolen (2016).

these actors, both interpersonal and institutional, is being affected by (un)intended consequences of the technology destroying the traditional modes and institutions of trust production. Thereby, the inquiry into how these systems produce impersonal trust takes on an additional role. The greater the ability of their technical features to instill confidence in users and convey a feeling of predictability⁶ of future events—independent of the traditional, institutional forms of impersonal trust—the stronger their impact is on societal trust in general. As we argue in this chapter, the impact of these technical features deserves extra scrutiny, if that impact is a growth of some form of trustlessness, which implies, in other words, the loss of (the need for) impersonal trust.

Despite the hype and the alleged revolutionary potential of these algorithmic digital technologies, they have yet to fully materialize. But, perhaps they don't need to. Their impact is measurable even if they don't fulfill their promises, and the disruption, as imagined, will not ultimately happen. The potential for damage is already there if individual consumers, citizens, public and private institutions embrace the unsubstantiated, never-to-be fulfilled claims about technological solutionism, the efficiency claims, the desirability of innovation at all costs, and the benefits of disruption.⁷ For this reason, in this chapter, we will not look as much into whether decentralized and automated socio-technical decision-making systems will effectively replace our existing institutional methods of producing trust. Instead, we focus on the following question: What kind of dynamics are set into motion by the introduction of these systems into societal trust production logics?

The next section will introduce the social labor of trust production. We will then line out how in the realm of identity management, pursuits towards trustlessness, enabled through decentralization, automation, and verification, may discard, what we refer to as, the 'second dimension' of societal trust, which encompasses the procedures, interactions, and safeguards to ensure not only trust, but also a sense of community.

2 The Social Labor of Trust Production

The initial source of trust in other people and institutions may be inherited, developed through repeated interactions and long-term experiences, or provided by external references. Thus, trust is typically earned through reliable behavior, positive experiences, or good reputation. Known actors garner trust through personal experiences, while trust in unknown actors may arise from the positive experiences of others, recommendations, or the safeguards given by external institutional intermediaries. At the other end, trust collapses quickly, if and when the other breaches that

⁶Predictability in this context is instilled through machines and artefacts, and is not coming from the act of trust, which is built from reliable, benevolent, and corrective behavior of societal actors.

⁷See Bodó and Janssen (2022) and Janssen et al. (2018).

trust. But what happens between the birth and death of trust? *What kind of investment, upkeep, and maintenance work are necessary to maintain and support both personal, interpersonal trust, and abstract, societal trust?*

For the purposes of this chapter, we do not discuss in detail how that process takes place in face-to-face, interpersonal trust relations, as this has already been extensively discussed.⁸ Instead, we will focus on the labor required to maintain the abstract and impersonal societal trust. Some of that trust is certainly the aggregate of interpersonal trust relations. If the members of a community generally trust each other in their micro-transactions, the overall, communal social trust should also be high. However, in modern societies, much of impersonal trust is produced by institutions, and through the interactions between citizens and institutions.⁹ Let us briefly sketch the societal labor necessary to maintain impersonal trust.

2.1 *The Social Labor of Impersonal Trust as a Shared and Social Resource*

We define impersonal trust as a societal, broadly accessible resource, that facilitates the co-existence, collaboration, and cooperation of strangers within a particular polity, usually the nation state.¹⁰ Such impersonal trust is produced by various public and private institutions and practices. As Zucker put it, when the mechanisms of trust production “*are reconstructed as intersubjective and as part of the “external world known in common” they can generalize beyond that transaction. This process of reconstruction has been called institutionalization: the process of redefining acts as exterior when intersubjective understanding causes them to be seen as part of the external world and objective when they are repeatable by others without changing the common understanding of the acts.*”¹¹ The source of such impersonal trust is external to the trust relation it supports. Such trust can rely on many sources, infrastructures, such as the media system or public education producing shared epistemic frameworks, the justice system producing reliable contracting and conflict resolution regimes, or the private sector providing risk management opportunities. Thus, impersonal trust can be considered a *shared* resource, as it should be accessible to everyone.

⁸ See Gambetta (1988), Goffman (1990), Greif (1989), and Hardin (2002). Let it suffice to say that potential trustees (i.e.: all of us) have to act in accordance with the formal and informal expectations of trustworthiness. This includes demonstrating one’s competence in certain tasks, performing a social role that demonstrates reliability, benevolence, care for others near and far, and acting with general integrity. Good social standing is a synonym for trustworthiness, and that social standing needs to be upheld both in the concrete, everyday social, economic interactions, as well as beyond them, in face of general social entropy and forgetfulness.

⁹ See Zucker (1985).

¹⁰ See Giddens (1990), Misztal (1996), and Sztompka (1999).

¹¹ See Zucker (1985).

There have been attempts to quantify the amount of resources spent on the production of such impersonal trust.¹² The study selected a number of professions in the US economy, such as managers (in business, finance, education, healthcare), service occupations (food preparation, care, etc.), sales, natural resource management, and transportation, and estimated how much time each profession spends with upholding trust. They found that around one third of the employment in the US is in one way or another related to upholding trustful economic relationships. This study frames this finding as the “cost of trust” and goes on to argue that at least some of this cost can be saved by relying on blockchains, smart contracts, and other algorithmic systems to produce the same trust, but cheaper. For the authors, trust production is first and foremost an economic question centered around costs and efficiency. However, reducing complex social, cultural phenomena to the question of economic efficiency robs them of all their non-economic functions. Fast food restaurants may be more efficient than cooking for friends, or maintaining a food service for the homeless, but it is also obvious what the more efficient alternative lacks.

The economic efficiency framing dis-embeds the practice it is applied to from the social relations it otherwise is embedded into.¹³ This also applies to how we see the labor spent on producing trust in society. The efficiency approach removes many of those considerations that socially embedded and controlled trust production practices embody and follow, such as care, maintenance of social relations, and ultimately the *maintenance and reproduction* of interpersonal and impersonal trust. Hence, impersonal trust is also a *social* resource, implying that economic and quantifiable considerations are unable to capture the multifaceted and abstract nature of trust.

2.2 *The Social Labor of Trust as a Shared Responsibility*

Since centuries, public institutions are the primary producers of impersonal trust, and the social labor of trust production takes place in the frameworks they offer. Schools don’t just teach children to read and write, they produce trust through diversity in the classroom, and creating the shared epistemic frameworks above and beyond individual, and group-based (racial, religious, linguistic) differences. Public service media doesn’t simply provide news and entertainment. While doing so, it’s mandate is to provide a continuously and freely accessible baseline in the cacophony of opinions, partisan debates, polarized info-bubbles. Likewise, a transparent, accountable, unbiased public administration doesn’t just manage communal resources, but while doing so, it allows trust to develop between those who rely on and contribute to those resources. Let’s take this latter trust production logic as an example to see in more detail how their trust production function may be affected

¹² See Davidson et al. (2018).

¹³ See Polanyi (1944).

by external pressures which don't take into account the disruption of the indirect trust related functions of public institutions.

The fate of the once lauded, but now dwindling New Public Management (NPM) approach highlights recent changes in the public governance and maintenance of trust. By introducing NPM, governments started to treat citizens as customers through mimicking private sector management models and leaning on market- and profit-oriented ideals, such as efficiency and performance measurements.¹⁴ NPM seeks to enhance public service delivery by promoting decentralization, competition, and customer-centric service provision. On the surface these goals are commendable, but the effect they had on trust relations only became evident long after their introduction. See, for example, the strategy brief on trust in the government by the City of Amsterdam, signed by the Mayor and the Secretary of Amsterdam:

In the 1980s, government organizations were seen as too cumbersome, bureaucratic, and slow. The government had to be 'run' like a business. Effective and efficient: achieving goals with as few people and resources as possible. Values such as humanity, fairness and openness became less important. The human dimension disappeared to the background. Citizens were seen as consumers and that had an effect on how governments treated citizens. Namely, as individuals who are mainly concerned with their own self-interest, people who cannot be trusted and therefore must be checked. If we want to increase residents' confidence in the municipality, we must take the first step and embrace a different, more positive view of humanity. Where we think and act as a starting point assume that Amsterdam residents are of good will. Trust instead of suspicion is the basis to act as a reliable government that primarily serves the citizen.¹⁵

This problem statement highlights the complexity of the role of public institutions as sources of impersonal trust. On the one hand, their mandate is to maintain trust in a community. For that, however, they need to be trusted by their constituents: the community they serve. Yet, this trust collapses, if the citizens see them as only focusing on efficiency, or in the name of the NPM ethos, they try to base their activities on control rather than trust.

Private institutions, such as banks, credit rating agencies, PR, marketing, and communications agencies also produce societal trust, either as their main area of activities, or as a byproduct of providing trust-requiring services. Unlike public institutions, their focus is not first and foremost to produce societally optimal levels and forms of trust. Instead, they must balance their economic interests, shareholder value, profits, market dominance with the kind of trust they are ultimately producing. Though the assumption of neoliberal economic theory is that economic self-interest, coupled with market competition would ultimately lead to trustworthy behavior in general, and in the case of trust producers, a steady and reliable supply of trust as a product, this does not always work in practice. The 2008 financial crisis was the result of financial institutions (investment banks, credit rating agencies)

¹⁴ See Ferlie et al. (1996).

¹⁵ Halsema and Teesink (2022).

prioritizing short-term profits over their long-term mandate¹⁶ to produce trust in economic relations, despite them being also heavily regulated to ensure their trustworthy behavior.

This conflict of interest is greatly aggravated in the case of technology companies in the business of trust production, where platforms, search engines, communication services, dating apps, self-driving car companies are repeatedly caught breaching their users' trust by selling their data, eavesdropping their conversations, feeding them misinformation, or putting them in physical danger, because profitability concerns seem to be more important than being trustworthy in all those relations they service.

2.3 *Algorithmic Technologies and the Social Labor of Trust*

The amount and nature of societal trust is the outcome of a constantly shifting balance between different trust production approaches: trust produced in interpersonal relations, trust produced by public institutions, and by private institutions. For example, if trust in public institutions is dwindling, which means that citizens rely less on the impersonal trust safeguards produced by the government, they may rely more on interpersonal trust relations.¹⁷ Consumers may be happy to replace some of the trust they feel towards other humans with trust produced by private actors, when, for example, they let their car drive itself, rather than grabbing the wheel themselves. Public institutions are increasingly outsourcing their trust production functions to private entities. The closure of schools during the COVID pandemic, and the increased reliance of private education technologies (from Zoom and Google classroom to various learning platforms) outsourced many functions of the school (instruction, practice, assessment) to these technological actors. There is also a constant shift between various private trust producers, such as the one between traditional banks and neobanks,¹⁸ or between traditional news organizations and social media.

Some of these shifts are less controversial than others. At the time, the shift from Encyclopedia Britannica to Wikipedia as a source of trustworthy and trusted information seemed outrageous, by now we know that ultimately there is little difference between the two in terms of the level of editorial control over individual articles, or the selection mechanisms producing the editors.¹⁹ Other shifts are way more controversial. Replacing seasoned journalists and newsrooms with AI and social media

¹⁶ In this sense, a mandate is not to be understood as a legal mandate, but for financial institutions to function properly in the financial system, trust is the very foundation. Without trust, customers would not deposit their money, seek financial advice, or use banking services.

¹⁷ See Fukuyama (1996).

¹⁸ See Ferrari (2022).

¹⁹ See the (self-)evaluation of Wikipedia on trustworthiness criteria: https://en.wikipedia.org/wiki/Wikipedia:Evaluating_Wikipedia_as_an_encyclopedia, its complex administrative structure:

may save costs for news producers or consumers but apparently results in a breakdown of some of the fundamental shared epistemic frameworks, from science to politics.²⁰ Replacing human social workers with algorithmic decisions can harm everyone: citizens, whose needs are denied; public institutions, who are reinforcing rather than lessening inequalities through their algorithmic tools; and the rest of society, outside of these relationships.²¹ The algorithmization of the public sector, when it fails, leads to a breakdown of social trust relations between citizens and institutions; and through the growing distrust in the institutions tasked to produce impersonal trust, it can end with a general breakdown of societal trust.²² Finally, the responsabilization of the citizen under the neoliberal state shifts responsibilities from public institutions to individuals and impersonal economic processes,²³ also in terms of trust production. The neoliberal subject is increasingly left to their own devices to find an answer to the question whether the unknown other is trustworthy, and to prove their own trustworthiness towards the others.

Blockchain-based systems present a particularly nasty configuration of these issues. They are designed to serve an arbitrary, ad-hoc cloud of anonymous, individual subjects, under the assumption that no one *in the system* is trustworthy; the users, due to their disposable identities cannot form a community (the problem of so-called “on chain governance” is still unresolved, and the governance of communities—if they exist—happens at least in part separately from what is happening on the decentralized infrastructure); and as a result, there is neither a need, nor a possibility for social trust in the system. This assumption of course is incredibly corrosive, if the system is deployed in a setting where social trust is still present. First, such a system in and by itself, does not, because it cannot, contribute anything to the maintenance of that pre-existing trust. Second, it may actively destroy social trust by injecting the assumption of untrustworthiness into the community and replacing trust by control. Third, to the extent it actually replaces (the functions) of a public institution, it removes this institution’s trust-producing labor from the maintenance of social trust.

For any of these reasons, efforts that aim to disrupt and replace trusted societal middlemen with trustless, algorithmic, or confidence-based systems deserve extra scrutiny. Identity management systems, described in the next section provide a comprehensive case study to just do that.

<https://en.wikipedia.org/wiki/Wikipedia:Administration>, its overview of editorial control: https://en.wikipedia.org/wiki/Wikipedia:Editorial_oversight_and_control.

²⁰ See Laufer and Nissenbaum (2023).

²¹ See Eubanks (2017).

²² See Bodó and Janssen (2022).

²³ See Brown (2017) and Weigl et al. (2023).

3 Identity Management as a Case Study for Trust Labor

Identification, both digital and analogue, is essential for parties who do not know or trust each other, but want to, or need to collaborate, or simply co-exist. There are multiple ways how identity information can be managed, maintained, disclosed, each contributing differently to the social production of trust. Although more granular distinctions are possible, we will adopt a general approach in this chapter and lay out three of them: paper-based public identity management, and privatized, so-called federated models (Sect. 3.1), and decentralized identity management systems, also known as self-sovereign identity systems (Sect. 3.2). In the following we will compare these alternatives from the perspective of their trust related functions.

3.1 *Trust and Distrust in Public and Private Centralized Digital Identity Systems*

3.1.1 Public Identity Infrastructures

The history of identification, public administration, and the collection of personalized information dates back thousands of years. During this time, governments emerged as the primary identity providers to maintain social order, manage taxation, and deliver public services. As societies grew in complexity, governments took on the role of centralizing and regulating identity documentation, leading to the development of formal identity information, such as date of birth, address, or nationality.²⁴ Given that identity information is largely rooted in the confirmation of an individual's citizenship, most types of personal identification have thus remained under the authority of governments.²⁵ An instance of an established instrument for confirming citizenship are paper-based passports. Serving as government-issued and globally acknowledged documents, passports enable individuals to, for instance, verify their identity when crossing international borders, and thereby facilitate travel.²⁶ When not coupled with surveillance technologies (such as biometrics, facial recognition systems, or GPS tracking) or politically and ideologically motivated misuse, they are also a trust anchor for public authorities and other service providers. The personal identification enabled through these official documents plays a relevant role for individuals engaging with public administrations, but also for accessing services provided by commercial entities, such as opening a bank account.

However, using physical documents for identity proofing can come with security risks, including the potential for breaches, identity theft, personal data misuse, and

²⁴ See Lips (2019).

²⁵ See Wihlborg (2013).

²⁶ See van Dijck and Jacobs (2020).

compromised privacy.²⁷ The digitization of identification systems signified a move toward fostering a more transparent and trustworthy information society and a step toward modernizing administrative and bureaucratic processes. This, however, introduced new risks and challenges to the public administration of identities, such as (growing threats of) cybercrimes, and a subsequently enhanced need to safeguarding users' privacy and security.

3.1.2 Private Identity Providers

With the advent of Facebook in the early 2000s, and various commercial services requiring user identification online, the traditional perception of identity management as a prerogative preserved for public authorities was substantially weakened. In the private sector, a digital identity management approach emerged, known as the 'federated' identity management system. In this system, private companies act as central trust authorities, enabling users to verify their identity through a single-sign-on process on the platform where they first registered their accounts. The system requires all involved parties to have a mutual sense of trust, since it is based on an arrangement between multiple domains that enables users to use the same identification data to access different networks. Hence, for commercial services, major private trust mediators like Google or Meta started to offer convenient identity management systems that store users' access credentials. Users 'only' need to establish a single account with one of these digital intermediaries, or identity providers in this case, to gain access to a diverse array of portals and services.

3.1.3 Trust Related Issues with Centralized Identity Management Systems

Both public and private digital identity management systems are in effect centralized repositories of personal data. This gives rise to issues of privacy, security, surveillance, and misuse, as the following examples testify.

Centralized data repositories of identity information are vulnerable to security breaches and data leaks, in some cases affecting billions of users, exemplified by cases such as those involving Yahoo!, LinkedIn, and Meta. Private actors are also prime targets for state surveillance, such as those revealed by Edward Snowden in 2013.²⁸ Classified documents exposed details about an electronic mass surveillance data mining initiative led by the NSA, with the collaboration of major tech companies such as Microsoft, Yahoo!, Google, Facebook, YouTube, and Apple, aimed at acquiring real-time data about US citizens. Private identity providers are also sensitive to how well, and by what values they are managed. Elon Musk's 2022 takeover

²⁷ Ibid.

²⁸ See Königs (2022).

of Twitter illustrates how the commercialization of the platform's user identity verification system, and the introduction of a subscription-based system can lead to the proliferation of fake accounts with verification checks, enabling the spread of misinformation and undermining the original purpose of the 'blue check' as a seal of authenticity of the account holder.

Despite all these concerns, users seem to have confidence in these systems to manage their (digital) identities. In case of the public systems, they have little choice. What is more surprising is their continued use of those private identity management systems which already proved themselves to be untrustworthy.

The reasons behind users' continued reliance on untrustworthy identity management systems can vary. Some users believe that they can manage and take safeguards against security and privacy risks, leading them to continue using the services of digital trust mediators with a strong belief in making informed decisions.²⁹ Digital service providers also do a lot to appear as trustworthy. Instances of such practices are (1) a strong focus on user experience and user-centric features, (2) some investment in internal governance mechanisms to increase trustworthiness, (3) relying on external signals and attestation of trustworthiness, (4) inviting regulation, or (5) building powerful (meta) narratives of trustworthiness, often without, or directly contradicting real world facts, and experiences.

These trust building practices, however, are quite different from actual, verifiable *trustworthiness*. Governments and private technology companies can only earn trustworthiness by being reliable and consistent in their behavior. Otherwise, these practices only produce misplaced trust: convincing users to trust something which isn't actually *worth* their trust. While trust-building actions can certainly enhance perceived, but possibly misguided trustworthiness, building and maintaining actual trustworthiness requires an often much more costly ongoing commitment and integrity.

The difference between investing in trustworthiness perceptions, and investing in actual, verifiable trustworthiness is what differentiates between the lack, and the presence of the social labor of trust production. The former ultimately decreases social trust by investing in deception, the latter actually contributes the social trust by investing in trustworthy trust producers.

The social labor of trust production in this latter case can take many forms. Trustworthiness of both public and private digital intermediaries in identity systems can be gained through, for example, user data and privacy protection through encrypted communication, compliance with data protection regulations, and the prioritization and demonstrated commitment to the proper and ethical use of any collected user information, at the cost of monetary gains, as well as other commercial and political interests.

²⁹ See Bodó et al. (2023).

3.2 *Trustlessness in Decentralized Digital Identity Management Systems*

Historically, institutional trust has experienced ups and downs. The last decades witnessed several downturns, marked by trust-shattering events invoked by political, digital and also financial actors.³⁰ These socio-economic conditions prompted technological responses that envisioned a disruption of traditional governance structures.³¹ Decentralized technologies like blockchain constituted an alternative governance model grounded in cryptographic code rather than contracts and legal frameworks, and verification mechanisms rather than institutional trust.³² The foundational technology of blockchain initially emerged within the cyberlibertarian realm of the internet, marked by a deep skepticism towards both public and private institutions.³³

The sphere of digital identity management was not left untouched, as engineers and a small circle of identity management experts called for a decentralized infrastructure as an approach to solve the trust issues provoked by centralized actors. With blockchain gaining prominence as a hype technology, a new libertarian blockchain-based digital identity paradigm touted as ‘self-sovereign identity’ (SSI) emerged. Taking into account the conceptual analysis of a variety of scholars studying this concept,³⁴ self-sovereignty can be understood as a concept in which data subjects exert control over their data and can determine how data are collected, processed, stored, and shared.

According to the SSI philosophy, safeguarding user privacy and security can thus only be achieved by removing any centralized entity involved in the exchange of user data. Essentially, data control should be decentralized, and this could be accomplished on a technical level through public key cryptography. Public key cryptography operates with key pairs, comprising a public key and a corresponding private key. The ownership of the key pair can be mathematically proven without disclosing the private key itself. Public key cryptography enables users to generate their own identifiers known as decentralized identifiers (DIDs), which facilitate encrypted communication between users and service providers without the necessity of a central authority for registering or revoking the identifiers. Similarly, verifiable credentials are a frequently observed technical feature of SSI. The validity of the credentials can be assessed without the involvement of the credential issuer and simply by controlling the digital signature of the issuer and a public revocation registry. For this purpose, a distributed ledger acting as a single source of truth regarding public information on credential issuers is recommended.³⁵ Users can then store DIDs,

³⁰ See Earle (2009), Levi and Stoker (2000) and Werbach (2018)

³¹ See Davidson et al. (2016).

³² See Werbach (2017).

³³ See Golumbia (2016), Karlstrøm (2014), and Reijers and Coeckelbergh (2018).

³⁴ See Couture and Toupin (2019), Herian (2020), Ishmaev (2021), and Weigl et al. (2023).

³⁵ See Mühle et al. (2018).

associated cryptographic keys and credentials on their devices or in the cloud, in a so-called ‘digital wallet’. For enhanced user privacy, computer scientists, cryptography researchers in particular, explored cryptographic methods for the selective disclosure of identity credentials. The assumption that, in general, centralized institutions would typically ask for more data than necessary, led to the integration of zero-knowledge proofs (ZKPs) as tools to minimize information exposure during a verifiable presentation and reveal only the attributes relevant to the specific purpose. ZKPs would also avoid the disclosure of an associated identifier, such as the public key of the issuer’s signature.³⁶

Thus, in order to verify someone’s claim or identity, the holder or user of the information that needs to be verified can provide verifiable credentials, that is cryptographic proofs, generated from their wallet to the verifier. These proofs contain information that was attested by an issuer beforehand. Attestation in SSI systems is enabled by the issuance of verifiable credentials, which are digitally signed by the issuer, and thereby assure the accuracy and authenticity of the information. Verifiers can independently check for the veracity and validity of these credentials by checking the cryptographic signatures of the issuer’s public keys, without the involvement of a centralized entity or authority.

Essentially, the technical characteristics of SSI are designed to bridge the high level of distrust in centralized entities and seek to grant individuals greater power and control over their personal data. From a normative point of view, however, it is worthwhile to critically scrutinize the moral and ethical feasibility of decentralized data control, and the replacement of digital intermediaries and institutions and their trustworthiness. Two main points within the context of trust disruption merit particular attention.

The first pertains to the *infrastructural* challenge stemming from what we can delineate as the ‘*ledger society*’.³⁷ The second involves an *individual-level* concern arising from the strong pursuit of *user empowerment* and control. The ledger society is characterized by great expectations and hopes regarding blockchain technology to solve societal issues. SSI, in this context, is “an idea that arguably assumes the retreat or abject failure of institutions, including intersubjective trust”.³⁸ Thereby, SSI does not necessarily seek to recalibrate or enhance social cohesion or political integrity. Instead, its objective lies in disrupting trust and establishing verification mechanisms and control as the bedrock for societal transactions encompassing both impersonal interactions and institutional relationships.³⁹ Secondly, as a user-centric identity management model, SSI aims to empower users with increased control and ownership of their personal information. However, the proponents of SSI, evidenced by the case of digital COVID-19 immunity passports, neglected ethical problems

³⁶ See Sedlmeir et al. (2022).

³⁷ See Herian (2020).

³⁸ Herian (2020), p. 157.

³⁹ See Bodó (2021a).

such as social stratification, discrimination, or deliberate self-infection.⁴⁰ SSI can thus be described as an economic model that establishes self-constituted markets for credentials and identity verification. Ideally, users and consumers feel autonomous and sovereign as they control the methods and avenues of data exploitation, practices that, however, inherently remain unchanged.⁴¹

These two dynamics on the infrastructural and the individual level eliminate the need for trust by establishing an ecosystem that is characterized by transparency, immutability, and user autonomy, and thereby eliminating the reliance on trusted intermediaries.

4 A Lost Sense of Community

4.1 *Trust Production in Identity Management*

As mentioned above, at a more abstract level, we can discern three distinct approaches to managing digital identity (public, private, and decentralized systems) and the underlying production of impersonal trust. However, to begin with, it is important to acknowledge that next to any public administrative function, such as identity management, trust is needed for a democratic state to operate in the first place. Well-placed trust is feeding into the legitimacy a state needs to function on behalf of its citizens. For public identity management in democratic societies, trust production relies on democratic control, the rule of law, and the accountability and legitimacy of public institutions. In turn, trustworthy identity management in the public realm is important as it allows the justice system to operate, it allows public registries to be trustworthy, and it allows public administration to fulfill its tasks to develop and implement just, equitable, accountable policies.

In the privatized or federated models, identity information is managed by private, profit-driven companies. They operate in a different territory as public institutions, and they are able to provide trust safeguards for interactions across jurisdictional boundaries at scale. How they do that, however, is only in part defined by the needs of the users and the specific interactions they mediate. Of course, on a technical level, they need to manage identity information in a safe, secure, and reliable manner, by protecting users' personal data and privacy, using encryption, and complying with data protection laws, as well as integrating multi-factor authentication or privacy-enhancing technologies. Yet, private actors have their own incentives: balancing safety and legal compliance against costs and following economic incentives which often result in identity information being used for purposes that benefit the identity provider and not always the users. They also don't have societal concerns in mind when they decide whom to issue a federated identity, what kind of compliance

⁴⁰ See Halpin (2020).

⁴¹ See Herian (2020).

they enforce, when and how they monitor bad behavior, and when to deny authentication to services. Their practices, norms, and rules may be oblivious to, or in some cases, in direct conflict with the purposes and goals of public identity providers.

If these practices are not effective enough, or if institutions are not perceived as trustworthy anymore, a third option to govern digital identities emerges; the decentralized identity paradigm. Decentralized identity systems aim to empower individuals to act autonomously without the interference of institutional middlemen. The trust placed in any central actor is minimized, despite the consequence of disconnecting from the broader societal system. In the evolving landscape of digital identity management, the desire to reduce the influence of untrustworthy federated and public intermediaries, particularly in the era of datafication, is a recurring theme. The incentive behind this shift is clear—a response to the numerous instances where trust has been betrayed through security breaches, personal data misuse for behavioral manipulation, microtargeting, data-driven profiling, mismanagement, and through political or ideological differences. All of that accumulates and results in institutional mistrust. The rationale behind embracing a trustless approach appears comprehensible. However, the notion of entirely removing intermediaries for digital identification and verification purposes raises several questions when analyzing trust from a sociological and philosophical perspective. This is because the function of trust goes beyond its role in facilitating and economizing interactions.

4.2 Two Sides of Two Different Coins: Technical and Societal Trust Mechanisms

In this section, we will contrast some of the self-sovereign identities' (SSI) technical features with the relevance of societal trust-building mechanisms. SSI, on a general technical level, is characterized by individual empowerment through enhancing privacy and anonymity, and the detachment from centralized systems. The infrastructure needs to be a decentralized network, such as a public or private blockchain or any decentralized public key infrastructure.⁴² Embedding identity verification (of citizens) in a trustless network comes with societal consequences, especially when considering the role trust plays in various interpersonal and institutional interactions in social and political life.

⁴² Although recent implementations of SSI systems have moved away from extensive blockchain integration, some still argue blockchains could be useful for hosting the public key infrastructure for certificate issuers and public registries that contain information on the revocation status of credentials. Regardless of the implementation of a blockchain, the SSI system always exercises some degree of decentralization with the usage of digital wallets operating on local devices and containing users' identity data.

4.2.1 The Disruption of Social Capital

First, for several reasons, life becomes more manageable in a community equipped with *social capital*. This term, as defined by Putnam, refers to “connections among individuals—social networks and the norms of reciprocity and trustworthiness that arise from them”.⁴³ Social networks cultivate norms of generalized reciprocity, enabling coordination and communication, enhancing reputations, and ultimately providing a means to resolve collective action dilemmas. Taking the idea of self-sovereign identities further will make social capital difficult to accumulate. Self-sovereignty enforces individual control and autonomy over personal data.

In a society where identity is managed by individuals through decentralized systems, the traditional bonds of trust and reciprocity, which are essential for the formation of social capital, become hardcoded as obligations and crypto-economic incentives into an infrastructure layer. Self-sovereign identity infrastructures may or may not be the future of many social, economic interactions and trust relations in the future, but in some way or another, the affordances of the technology will shape those relations. A technology built on the assumption of anonymity, zero-knowledge proofs, distrust, and libertarian ideas of control and individuality will frame the interactions where this technology is put in use. Whether in interpersonal relations, or vis-à-vis public institutions, a technology of distrust will shape and constrain the interaction. It is hard to imagine how such a framework would be conducive to the emergence of collective trust necessary for social cohesion. In addition, rules, hardcoded into the infrastructure layer make it difficult for a community to adjust its own norms to changing circumstances because such capacity is now tied to the restricted technological domain of protocol development.

Another aspect is how such technologies can contribute to the development of shared epistemic frameworks. Lynne Zucker argues that trust emerges from a complex set of background expectations, the attitudes of daily life, reciprocal perspectives, and socially warranted knowledge.⁴⁴ This is a two-way interaction. Common background expectations produce trust, and it is the social labor of trust production which to a large extent forms shared background expectations. These expectations are constantly challenged and reassessed according to fast-changing global and societal circumstances. In the epistemological framework of SSI-based trustlessness, this process, the formation and modification of background expectations, is reduced to the mere impression of confidence that the system works. The interactions individuals build through such systems also do not allow for the production of background expectations, or for an update thereof, as SSI does not produce trust, it produces verification mechanisms, such as cryptographic protocols and attestations, due to distrust. With each interaction, we calibrate what we can expect from each other and the environment. In the absence of a functional trust system based on common expectations, society may fragment into tribal trust networks. The danger

⁴³ See Fukuyama (1996) and Putnam (2001), p. 19.

⁴⁴ See Zucker (1985).

of such fragmentation becomes apparent as societal systems lose their coherence. This results in dysfunctional interactions, hindering the development of broader, interconnected communities.

4.2.2 The Disruption of Civic Culture

Secondly, trust is also a central element of *civic culture*. As Sztompka argues, “its presence is an indispensable precondition of a viable political system”.⁴⁵ The foundation of the political system rests on active participation, where the crucial role of engagement lies in incorporating practices that allow everyone to witness and contribute to the system’s trustworthiness—often by participating in practices of distrust: control, oversight, verification. Consequently, trustworthiness becomes an experience of engagement. From a certain perspective, distrust can be seen as beneficial as it implies a sense of guardianship, wherein we scrutinize those who are in positions of authority.

In SSI systems, zero-knowledge proofs (ZKPs) are optimized to empower users by enhancing anonymity through privacy-enhancing features. Identities are verified through cryptographic protocols (cryptographic hash functions and mathematical proofs), which allow the holder of identity data to demonstrate knowledge of a secret to a verifier without revealing that information itself. While ZKP can allow for identity verification without oversharing specific information, this goal comes at the expense of being able to participate in the production of collective societal goods, such as trustworthiness, which relies on some levels of visibility and legibility.⁴⁶ For instance, in the context of law enforcement, excessive anonymity resulting from their use might impede investigations into criminal activities.⁴⁷ Healthcare institutions may face challenges in efficiently managing public health crises if they cannot access anonymized data for epidemiological analysis.⁴⁸ Moreover, since ZKPs have to date no legal development under EU regulation, there is a risk of impeding regulatory efforts. Hence, while ZKPs provide a cryptographic means to validate information without revealing it and constitute an essential feature in a zero-trust environment, the use of such systems contributes little to the operation and maintenance of a trustworthy system itself. If we delegate this responsibility to an algorithm of trustlessness, we forfeit the ability to uphold trust through vigilant skepticism. SSI is unable to provide this process because this task has been rendered irrelevant by delegating it to the protocol level.

⁴⁵ Sztompka (1999), p. 14; and see also Almond and Verba (1963).

⁴⁶ See Golman et al. (2017), Kahneman and Tversky (1979), Pozen (2018), and Prat (2005).

⁴⁷ See Garland (2001), Hert and Gutwirth (2006), and Jardine (2015).

⁴⁸ See Abouelmehdi et al. (2018) and Gille and Brall (2021).

4.2.3 The Disruption of Societal Interactions

Thirdly, trust is also an important dimension of *civil society interactions* in general. The existence of a functioning community of citizens, dedicated to political authority, relies on both mutual trust among individuals and a trusting relationship with public institutions. Institutions thereby also serve as the catalysts for a continuous calibration process between individuals. This calibration process involves language, concepts, and the establishment of a shared understanding. Removing or neglecting this capacity, which is fundamental to trust, disrupts the process of building a common understanding. Emerging decentralized systems like self-sovereign identity further underscore the effect of postmaterialism on societal trust, whereby individual values shift from physical and economic principles to values like autonomy and self-expression.⁴⁹ Thus, while typically supportive of democratic ideals, postmaterialism also embodies perspectives and actions that challenge the establishment, along with growing discontent with authority within contemporary democracies. This resonates particularly with the perspective of blockchain evangelists who argue that the technology has the potential to provide internationally recognized self-sovereign legal identities to everyone. This could, in turn, empower individuals to establish virtual communities in cyberspace, dedicated to political decision-making and the formulation of laws, surpassing the confines of (national) institutions.⁵⁰ It is plausible that individuals supporting such views largely contest authority and exhibit a heightened pursuit of trustlessness.⁵¹

In conclusion, while the desire to reduce the power of trust intermediaries in the evolution of digital identity management is understandable, a nuanced approach is necessary. Trust is not just a procedural safeguard but a vital element in civic culture and civil societies in general. Institutions and collective practices play a crucial role in this process. Various forms of identity management encourage specific forms of trust production and societal reproduction and discourage others. The choice between the different approaches thus has consequences beyond identification and verification of different aspects of our identities in narrowly defined settings. Replacing public identity infrastructures with self-sovereign ones means abolishing certain forms of trust labor, and thereby risking societal fragmentation and the breakdown of cohesive communities.

It can finally also be argued that if SSI was really able to foster trust by making issuers, verifiers, and users more trustworthy, a significant contradiction arises. The paradox lies in the potential resistance from entities, such as private verifiers, who may have vested interests in monetizing data or exploiting extensive data collection practices. In this scenario, the implementation of mechanisms to verify credentials and empower users with ZKPs and control over their digital identities could face pushback from those accustomed to traditional data-centric models. Furthermore,

⁴⁹ See Catterberg and Moreno (2006) and Inglehart (1990).

⁵⁰ See Orgad (2018).

⁵¹ See Catterberg and Moreno (2006).

when governments or supranational actors, like the EU, adopt SSI, there exists a possibility that these entities might embrace SSI simply to demonstrate a commitment to principles of transparency and individual empowerment, which they are already perceived to uphold. On the other side of the coin, if the SSI technology lived up to its promise of state ‘disempowerment’ and individual autonomy, would untrustworthy surveillance and authoritarian states adopt SSI and fully relinquish the control to citizens, just to become trustworthy?

5 Conclusion

In the realm of digital identity management, small circles of experts have advocated for a decentralized infrastructure as a solution to address the trust issues arising from centralized actors. With the rise of blockchain as a ‘trustless’ technology, a novel libertarian digital identity paradigm known as SSI has surfaced. Decentralized systems like SSI alter the form, content, shape, and amount of impersonal trust in often unforeseen and unforeseeable ways. The aggregate impact of the changes is hard to assess. Yet, in this chapter, we asked what kind of dynamics are set into motion by the introduction of trustless systems into societal trust production logics?

We posit that societal, impersonal trust facilitates the co-existence among strangers within a specific political entity. This form of trust is generated through diverse public and private institutions and practices as laid out in Sect. 2. From this, we formulated some normative goals, when it comes to the social production of impersonal trust. First, impersonal trust is a *shared* social resource. This means that ideally, it is equally accessible to all members of the community, without discrimination. Second, impersonal trust is a shared *social* resource. This means that the nature, form, and amount of that trust should be defined by social, rather than purely or predominantly economic considerations and/or technological limitations. Private trust producers’ contribution to impersonal trust should also be assessed and ensured from the perspective of the community. And third, the production of impersonal trust is a shared *responsibility* of the community. This means that no member should face a barrier to, or completely excuse themselves from participating in the practices of trust production. Also, public institutions’ mandate should include this responsibility, both directly or through their regulatory and oversight functions vis-a-vis private parties.

In the case of SSI, two main elements of trust disruption clash with at least one of the normative goals above. The first revolves around the *infrastructural* disruption due to decentralization, which clashes with the second (*social*) and third (*responsibility*) normative goal. In this context, the objective of SSI is to disrupt trust and establish verification mechanisms and control as the fundamental basis for societal transactions, encompassing both impersonal interactions and institutional relationships. This removes trust as a *social* resource and as a shared *responsibility*. Specifically, digital wallets and DIDs eliminate the need to trust central entities by allowing users to control their data and communicate directly by leveraging

cryptography instead of relying on intermediaries. Distributed ledger technology eliminates the need to trust issuing entities, allowing to verify issuers' signatures for the integrity of credentials. Verifiable credentials eliminate the need to trust users, by making credentials tamper-evident and cryptographically verifiable, reducing the risk of forgery. ZKPs eliminate the need to trust verifiers, by minimizing the amount of personal data exchanged during verification processes and the potential for misuse of information.

The second clash pertains to *individual*-level concerns stemming from the intense pursuit of user empowerment and control. This user empowerment, from the perspective of trust labor, however, seems one-sided and superficial, and conflicts with the first (*shared*) and the second (*social*) normative goal. Some users may be empowered in a technical sense if they control their own identity information, but this does not translate into bargaining power vis-à-vis private entities trying to extract as much information from them as possible, and certainly results in disempowerment when it comes to participating in the social practices of trust production and trust development. Moreover, empowering users comes with attributing a significant responsibility to them. This responsibility also comes with accountability, but through the distribution of power and decision-making capacity, accountability can no longer be exercised by a protective central entity. Consequently, how is that entity supposed to be trustworthy when it cannot be held accountable? In essence, the pursuit of user empowerment requires careful consideration of the broader societal context, acknowledging the inherent challenges and responsibilities associated with empowering individuals within existing power structures. Simply providing users with technological tools does not automatically address the underlying power imbalances. To conclude, it remains unlikely that decentralized technologies will ever fully substitute the sense of community and belonging for individuals, even if these individuals harbor distrust towards societal institutions.

In this chapter, we used decentralized identity management as a case study for zero-trust technologies to show how, regardless of the numerous trust breaches, trust cannot be replaced by a techno-centric approach. It can, at best, make existing procedures more efficient and, if implemented correctly, sometimes more transparent. However, when sold under the premise of improving trust and contributing to a more trustworthy society, this does not fly. Though electronic identification is but one activity of governmental and commercial interactions with citizens and customers, it is of course also linked to the bigger picture of the democratic state and rule of law. This is because, electronic identification, if misused and exploited, easily lingers around infringements of data protection, privacy, and security, and therefore plays an important role in the connection between the state, its political system, and societal trust.

The right approach to address the perceived or actual untrustworthiness of institutional intermediaries is to be more engaged in the production of trust, instead of abandoning them to a trustless, protocol-based, self-sovereign alternative. The labor of individuals and communities to voice criticism and hold entities accountable is the means of advancing toward improvement. Ultimately, it should also not be left unsaid that a well-functioning society relies on the establishment of mutual

institutional trust. Institutions will find it hard to effectively offer good governance, services, assistance, security, and support to an anonymous society. In an era where alienation, social fragmentation, polarization, discrimination, and growing distrust are major societal issues, it seems strange to try to solve these challenges by moving to an identity infrastructure that is built upon the exact same premises, with no intention, or technical affordance to address or overcome them. Instead of disproportionately channeling efforts into replacing institutions and disrupting entire systems through trustless technologies, a more sustainable approach for the greater societal good would involve identifying untrustworthy individual actors and patterns, while concurrently directing resources towards rebuilding and regaining trust within the existing framework.

References

- Abouelmehdi K, Beni-Hessane A, Khaloufi H (2018) Big healthcare data: preserving security and privacy. *J Big Data* 5(1):1. <https://doi.org/10.1186/s40537-017-0110-7>
- Almond GA, Verba S (1963) *The civic culture: political attitudes and democracy in five nations*. Princeton University Press. <https://www.jstor.org/stable/j.ctt183pnr2>
- Beck R, Czepluch J, Lollike N, Malone S (2016) *Blockchain - the gateway to trust-free cryptographic transactions*
- Bijker WE, Hughes TP, Pinch TJ (1989) *The social construction of technological systems: new directions in the sociology and history of technology*. MIT Press
- Bodó B (2021a) Mediated trust: a theoretical framework to address the trustworthiness of technological trust mediators. *New Media Soc* 23(9):2668–2690. <https://doi.org/10.1177/1461444820939922>
- Bodó B (2021b) The commodification of trust. *SSRN Electron J*. <https://doi.org/10.2139/ssrn.3843707>
- Bodó B, Janssen H (2022) Maintaining trust in a technologized public sector. *Policy Soc* 41(3):414–429. <https://doi.org/10.1093/polsoc/puac019>
- Bodó B, Bene M, Boda Z (2023) Standing naked in the storm—European citizens’ trust in social media, users, information. *SSRN Scholarly Paper*. Rochester, NY, February 23, 2023. <https://doi.org/10.2139/ssrn.4368419>
- Botsman R (2017) *Who can you trust? How technology brought us together and why it might drive us apart*, 1st edn. Public Affairs, New York
- Brown W (2017) *Undoing the Demos: Neoliberalism’s Stealth Revolution*. First paperback edition, this Edition corrects errors in previous printings. Near Futures. Zone Books, New York
- Catterberg G, Moreno A (2006) The individual bases of political trust: trends in new and established democracies. *Int J Public Opin Res* 18(1):31–48. <https://doi.org/10.1093/ijpor/edh081>
- Christin A (2016) From daguerreotypes to algorithms: machines, expertise, and three forms of objectivity. *ACM SIGCAS Comput Soc* 46(1):27–32. <https://doi.org/10.1145/2908216.2908220>
- Couture S, Toupin S (2019) What does the notion of ‘sovereignty’ mean when referring to the digital? *New Media Soc* 21(10):2305–2322. <https://doi.org/10.1177/1461444819865984>
- Davidson S, De Filippi P, Potts J (2016) *Disrupting governance: the new institutional economics of distributed ledger technology*. *SSRN Scholarly Paper*. Rochester, NY, July 19, 2016. <https://doi.org/10.2139/ssrn.2811995>
- Davidson S, Novak M, Potts J (2018) The cost of trust: a pilot study. *J Br Blockchain Assoc* 1(2):1–7. [https://doi.org/10.31585/jbba-1-2-\(5\)2018](https://doi.org/10.31585/jbba-1-2-(5)2018)

- De Filippi P, Mannan M, Reijers W (2020) Blockchain as a confidence machine: the problem of trust & challenges of governance. *Technol Soc* 62:101284. <https://doi.org/10.1016/j.techsoc.2020.101284>
- Earle TC (2009) Trust, confidence, and the 2008 Global Financial Crisis. *Risk Anal* 29(6):785–792. <https://doi.org/10.1111/j.1539-6924.2009.01230.x>
- Eubanks V (2017) Automating inequality: how high-tech tools profile, police, and punish the poor. St. Martin's Press, New York, NY
- Ferlie E, Ashburner L, Fitzgerald L (1996) The new public management in action. Oxford University Press
- Ferrari MV (2022) The platformisation of digital payments: the fabrication of consumer interest in the EU FinTech Agenda. *Comput Law Secur Rev* 45:105687. <https://doi.org/10.1016/j.clsr.2022.105687>
- Fukuyama F (1996) Trust: the social virtues and the creation of prosperity. 1. Free Press paperback ed. A Free Press Paperbacks Book. Free Press, New York
- Gambetta D (1988) Can we trust trust. In: Gambetta D (ed) Trust: making and breaking cooperative relations. Basil Blackwell, Oxford, England, pp 213–237
- Garland D (2001) The culture of control: crime and social order in contemporary society. Oxford University Press, Oxford, New York
- Giddens A (1990) The consequences of modernity. Polity Press, Cambridge
- Gille F, Brall C (2021) Limits of data anonymity: lack of public awareness risks trust in health system activities. *Life Sci Soc Policy* 17(1):7. <https://doi.org/10.1186/s40504-021-00115-9>
- Goffman E (1990) The presentation of self in everyday life. 1. Anchor Books ed., rev. Ed. Anchor Books, New York
- Golman R, Hagmann D, Loewenstein G (2017) Information avoidance. *J Econ Lit* 55(1):96–135. <https://doi.org/10.1257/jel.20151245>
- Golumbia D (2016) The politics of bitcoin: software as right-wing extremism. University of Minnesota Press, Minneapolis
- Greif A (1989) Reputation and coalitions in medieval trade: evidence on the Maghribi traders. *J Econ Hist* 49(4):857–882. <https://doi.org/10.1017/S0022050700009475>
- Halpin H (2020) A critique of immunity passports and W3C decentralized identifiers. arXiv, November 30, 2020. <https://doi.org/10.48550/arXiv.2012.00136>
- Halsema F, Teesink P (2022) Werken aan meer vertrouwen tussen Amsterdammers en de gemeente. Raadsinformatiebrief. City of Amsterdam, Amsterdam. July 30, 2022. <https://openresearch.amsterdam.nl/page/89681/werken-aan-meer-vertrouwen-tussen-amsterdammers-en-de-gemeente>
- Hardin R (2002) Trust and Trustworthiness. The Russell Sage Foundation Series on Trust, vol 4. Russell Sage Foundation, New York
- Hawliczek F, Notheisen B, Teubner T (2018) The limits of trust-free systems: a literature review on blockchain technology and trust in the sharing economy. *Electron Commerce Res Appl* 29:50–63. <https://doi.org/10.1016/j.elerap.2018.03.005>
- Herian R (2020) Blockchain, GDPR, and fantasies of data sovereignty. *Law Innov Technol* 12(1):156–174
- Hert P, Gutwirth S (2006) Privacy, data protection and law enforcement. Opacity of the individual and transparency of power. Privacy, Data Protection and Law Enforcement, Opacity of the Individuals and Transparency of Power, Privacy and Criminal Law 18. <https://doi.org/10.1111/rdp.v18i100.6200>
- Hughes TP (1987) The evolution of large technological systems. In: Bijker WE, Pinch TJ (eds) The social construction of technological systems: new directions in the sociology and history of technology. MIT Press, Cambridge
- Inglehart R (1990) Culture shift in advanced industrial society. In: Culture shift in advanced industrial society. Princeton University Press, pp 1–2. <https://doi.org/10.1515/9780691186740-003>
- Ishmaev G (2021) Sovereignty, privacy, and ethics in blockchain-based identity management systems. *Ethics Inf Technol* 23(3):239–252. <https://doi.org/10.1007/s10676-020-09563-x>

- Janssen M, Rana NP, Slade EL, Dwivedi YK (2018) Trustworthiness of digital government services: deriving a comprehensive theory through interpretive structural modelling. *Public Manag Rev* 20(5):647–671. <https://doi.org/10.1080/14719037.2017.1305689>
- Jardine E (2015) The Dark Web Dilemma: Tor, Anonymity and Online Policing. SSRN Scholarly Paper. Rochester, NY, September 30, 2015. <https://doi.org/10.2139/ssrn.2667711>
- Kahneman D, Tversky A (1979) Prospect theory: an analysis of decision under risk. *Econometrica* 47(2):263–291
- Karlstrøm H (2014) Do libertarians dream of electric coins? The material embeddedness of bitcoin. *Distinktion: Scand J Soc Theory* 15:23–36. <https://doi.org/10.1080/1600910X.2013.870083>
- Keymolen ELO (2016) Trust on the line: a philosophical exploration of trust in the networked era. Erasmus University Rotterdam. <hdl.handle.net/1765/93210>
- Königs P (2022) Government surveillance, privacy, and legitimacy. *Philos Technol* 35(1):8. <https://doi.org/10.1007/s13347-022-00503-9>
- Laufer B, Nissenbaum H (2023) Algorithmic Displacement of Social Trust. 23-12 Knight First Amend. Inst. (blog), November 29, 2023. <http://knightcolumbia.org/content/algorithmic-displacement-of-social-trust>
- Levi M, Stoker L (2000) Political trust and trustworthiness. *Annu Rev Polit Sci* 3(1):475–507. <https://doi.org/10.1146/annurev.polisci.3.1.475>
- Lips M (2019) Digital government: managing public sector reform in the digital era, 1st edn. Routledge
- Misztal BA (1996) Trust in modern societies: the search for the bases of social order. Polity Press, Cambridge
- Mühle A, Grüner A, Gayvoronskaya T, Meinel C (2018) A survey on essential components of a self-sovereign identity. *Comput Sci Rev* 30:80–86. <https://doi.org/10.1016/j.cosrev.2018.10.002>
- Orgad L (2018) Cloud communities: the dawn of global citizenship? In: Bauböck R (ed) *Debating transformations of national citizenship*, IMISCOE Research Series. Springer International Publishing, Cham, pp 251–260. https://doi.org/10.1007/978-3-319-92719-0_46
- Polanyi K (1944) The great transformation: the political and economic origins of our time. Beacon Press
- Pozen D (2018) Transparency's Ideological Drift. SSRN Scholarly Paper. Rochester, NY. <https://papers.ssrn.com/abstract=3120807>
- Prat A (2005) The wrong kind of transparency. *Am Econ Rev* 95(3):862–877. <https://doi.org/10.1257/0002828054201297>
- Putnam RD (2001) *Bowling alone: the collapse and revival of American community*, 1st edn. Touchstone Books by Simon & Schuster, London
- Reijers W, Coeckelbergh M (2018) The blockchain as a narrative technology: investigating the social ontology and normative configurations of cryptocurrencies. *Philos Technol* 31(1):103–130. <https://doi.org/10.1007/s13347-016-0239-x>
- Sedlmeir J, Barbereau T, Huber J, Weigl L, Roth T (2022) Transition pathways towards design principles of self-sovereign identity. ICIS 2022 Proceedings, December 12, 2022. https://aisel.aisnet.org/icis2022/is_implement/is_implement/4
- Sztompka P (1999) *Trust: a sociological theory*. Cambridge University Press
- van Dijk J, Jacobs B (2020) Electronic identity services as sociotechnical and political-economic constructs. *New Media Soc* 22(5):896–914. <https://doi.org/10.1177/1461444819872537>
- Weigl L, Barbereau TJ, Fridgen G (2023) The construction of self-sovereign identity: extending the interpretive flexibility of technology towards institutions. *Gov Inf Q* 40(2)
- Werbach K (2017) Trust, but verify: why the blockchain needs the law. SSRN Scholarly Paper. Rochester, NY, August 1, 2017. <https://doi.org/10.2139/ssrn.2844409>
- Werbach K (2018) *The blockchain and the new architecture of trust*. Illustrated Edition. The MIT Press, Cambridge, London, England

Wihlborg E (2013) Secure Electronic Identification (eID) in the intersection of politics and technology. *Int J Electron Gov* 6:143–151. <https://doi.org/10.1504/IJEG.2013.058371>

Zucker LG (1985) Production of trust: institutional sources of economic structure, 1840 to 1920. In: Cummings LL, Staw B (eds) *Research in organizational behavior*. JAI Press, Greenwich, Conn

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.



Learning-Loop Bureaucracies: A Configuration Approach to AI, Law, and Organizations



Stavros Zouridis

Abstract From ancient times onwards government bureaucracies have used technology. Their operations shaped the technologies but, in turn, the technologies also reshaped the operations. The massive use of AI will again reshape the organizational processes of large-scale government bureaucracies. Since these processes usually follow the law an analysis of AI driven large-scale government bureaucracies should take into account how AI affects the law. This chapter builds on a configuration approach that connects technology, the basic characteristics of the organization, and law. The configuration approach is first applied to understand how street-level bureaucracies transformed into screen-level bureaucracies and then into system-level bureaucracies. In each stage of the development of large-scale government bureaucracies the technology, the characteristics of the organization, and law constitute a stable configuration. The use of AI is still in its early stages but it is possible to explore the stable configuration that AI may trigger. This configuration is labeled as a learning-loop bureaucracy in which the individual decisions (such as welfare benefits or construction permits) result from man-machine interaction in a continuous learning-loop cycle. In order to constitutionalize these learning loops we may expect a new type of legislation that regulates the learning processes instead of the individual decisions.

S. Zouridis (✉)

Department of Public Law and Governance, Tilburg Law School, Tilburg University,
Tilburg, Netherlands

e-mail: s.zouridis@tilburguniversity.edu

© The Author(s) 2025

J. Goossens et al. (eds.), *Public Governance and Emerging Technologies*,
https://doi.org/10.1007/978-3-031-84748-6_17

1 An Early Discovery: Law Mediates Between Technology and Organization

Government bureaucracies prepare and execute public policies.¹ Since time immemorial, these bureaucracies have used technologies to support their operations. In ancient Egypt, the pharaohs used the abacus (a system of beads) for administrative purposes like taxes while the Inca people developed the quipu (a system of knots) for storing information and supporting the administration. Governments have always relied on technology for their core processes such as administration and registration. In turn, it may be expected that technologies have also shaped the everyday operations of government bureaucracies. Nothing new here. Research indicates that the introduction of new technologies is not neutral with regard to the operations for which technologies are deployed.² While technologies are shaped by the operations for which they are used, in turn the technologies shape these operations.³ Only think about how cars changed the nature of police patrols or how television reshaped communication between government and citizens. Again, nothing new here, and artificial intelligence (AI) can in this respect be seen as just another step in an ongoing process. Given the massive development and use of AI-technologies in government, technology will reshape government bureaucracies once more.⁴ The key issue is how these bureaucracies will change because of AI. This chapter explores the plausible organizational transformations that AI may trigger in government bureaucracies. The question addressed in this chapter is how AI affects the operations of the large-scale bureaucracies that implement and enforce legislation (e.g. welfare benefits, traffic fines, construction and environmental permits, and so on).

In order to analyze how AI shapes and reshapes large-scale bureaucracies this chapter uses organizational configuration theory.⁵ Configuration theory refers to a school of thought in organization studies which conceives an organization as a stable configuration of characteristics. The configuration of these characteristics aligns with a particular organizational environment. The alignment between this organizational configuration and its environment produces stability for the organization. In the lifetime of an organization, these stable episodes alternate with shifts to a different configuration that matches with a different environment. The configuration approach builds on organizational contingency theories that predict organizational change as a consequence of changes in the environment of the organization.

Contingency theories predict that organizations only change if the organizational environment changes. Co-evolution of the organization and its environment

¹ See Weber (1922), Waldo (1948), Dunsire (1978) and Wilson (1989).

² See Winner (1980), van der Pot (1994) and Kelly (2010).

³ See Woodward (1958) and Thompson (1967).

⁴ See Bullock (2019) and Flügge et al. (2021).

⁵ Originally articulated by Chandler (1962). See also Mintzberg et al. (1998).

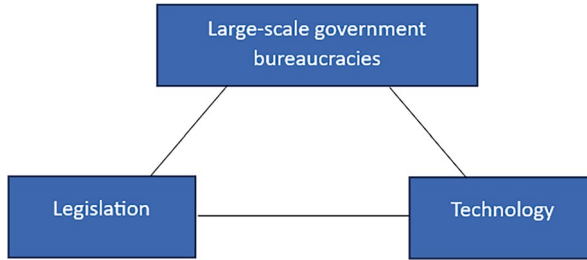


Fig. 1 The configuration of large-scale government bureaucracies, legislation, and technology

empirically appears to account only for a part of the organizational change observed.⁶ Sometimes organizations change while the environment does not change, so contingency theory then cannot account for the changes. Contingency theories also view the relation between the environment and the organization as rather one-sided. Empirical research shows that organizations however do not only depend on the environment but also actively shape the environment in which they operate.⁷ Configuration theory builds on contingency theory while addressing these flaws. In configuration theories organizations and their environment are seen as episodic configurations that change from time to time. Technological developments such as AI change the environment of large-scale government bureaucracies, triggering a shift in the entire configuration.

If applied to large-scale government bureaucracies, configuration theory shows at least three components that constitute a coherent configuration.⁸ Since these large-scale bureaucracies implement legislation, the design and the details of the legislation that is implemented are of major importance. Technological development is a second major component of the configuration that also includes these large-scale bureaucracies. In theory, thus the configuration of large-scale government bureaucracies that implement legislation encompasses the organization of these bureaucracies as well as the legislation that underlies their operations and the technologies these bureaucracies use for their operations. The configuration is depicted in Fig. 1.

Since the configuration approach predicts coherent configurations of the legislation, key features of the organization (large-scale bureaucracy), and the technology it may be expected that the rise of new technologies such as AI cause a shift in the configuration and trigger a new coherent configuration. To understand the expected configuration shift, this chapter first reflects on how the current configuration of technology, legislation, and large-scale bureaucracy has evolved. In Sect. 2, three successive configurations are sketched that have been observed in large-scale

⁶ See Mintzberg et al. (1998).

⁷ Ibid.

⁸ See Bovens and Zouridis (2002).

bureaucracies over the past five to seven decades.⁹ The configurations show that the design and the content of legislation co-evolved with the basic organizational principles of the large-scale bureaucracies that implement the legislation as well as the technologies deployed by these bureaucracies. Since AI has only recently been adopted in these organizations, it is too early to already predict what the new configuration will look like. Given the nature and characteristics of artificial intelligence, Sect. 3 aims to explore a plausible shift in the configuration of legislation and organization (bureaucracy) triggered by AI. While it is impossible to predict the future of these large-scale bureaucracies in detail, configuration theory enables us to draw some contours of a stable and coherent configuration in which AI is a principal component. The final section includes a brief discussion of the implications of the new configuration from a constitutional perspective.

2 Back in Time: From a Street-Level-Centered to a Production-Centered to an Information-Centered Configuration

Many large-scale government bureaucracies primarily implement policies and legislation.¹⁰ Tax authorities impose and collect taxes, welfare benefit organizations decide on applications for social security benefits, child allowance, and spatial planning bureaucracies decide on building permits. Large-scale law enforcement organizations monitor compliance, respond to non-compliance, deal with criminal offences, or provide residence permits or green cards. The application of legislation to individual situations is the common denominator of these large-scale bureaucracies. Traditionally the key operations of these organizations have been referred to as case processing or people processing by street-level bureaucrats.¹¹ Research in the Netherlands shows that ever since the 1960s these organizations have increasingly developed information and communication technologies (ICTs) to support their work.¹² Bovens and Zouridis have developed a framework to understand the changes in these organizations that result from the increasing use of ICTs.¹³ In this framework, the organizations successively changed from street-level bureaucracies to screen-level bureaucracies to system-level bureaucracies. These changes follow the different patterns of industrial organization already sketched by Woodward in the 1950s: from unit-based production to mass production to continuous process production.¹⁴ If applied to two large-scale bureaucracies in the Netherlands (the agency

⁹ See Zouridis (2000).

¹⁰ See Wilson (1989).

¹¹ See Prottas (1979), Lipsky (1980) and Maynard-Moody and Musheno (2003).

¹² See Zouridis (2000).

¹³ See Bovens and Zouridis (2002).

¹⁴ See Woodward (1958).

that processes study grants and the agency that processes traffic violations and fines) the configuration theory approach demonstrates the interplay between organization characteristics, technology, and legislation. Zouridis aimed to study the effect of technological changes on the large-scale bureaucracies.¹⁵ In order to see whether the same effects occur in different policy fields two policy fields were selected that represent different faces of government (welfare policies vis-à-vis security and law enforcement policies). These changes can be briefly described as follows.¹⁶

2.1 *Study Grants and Loans*

In the Netherlands, a system of scholarships for university students was already established early in the twentieth century. Only a small number of students then applied for a study grant and these applications were handled by the Ministry of Education. The officials responsible for the decisions on these applications were typical street-level bureaucrats as depicted by Lipsky.¹⁷ They knew the details of each student who applied, and they could assess the criteria that made students eligible for a study grant. If students were both gifted enough to study at a university and indigent because of the income and property of their parents, they could apply for a study grant. The officials assessed both the ability to study at a university and the financial circumstances of the students and their parents. If necessary, the officials could also contact the university to check whether the students with a grant worked hard enough to justify a study grant. Or to assess their circumstances, they invited the student to the Ministry. The amount of the grant was decided by the officials, and it was contingent upon the circumstances of the student. The legislation underlying this system was quite brief and allowed for quite some discretion both with regard to the circumstances and with regard to the amount of the study grant. Since study grants were seen as a stimulus given by the government, the legislation did not grant a right but only a rough sketch of the government's discretion. The philosophy underpinning the system was that just as students were free to apply for a grant or a university program the government was free to decide whether or not to grant a scholarship. Sometimes the income of the parents was merely a question of estimation for the officials because as small shopkeepers they did not maintain accurate financial accounts. Students from an urban environment also were assumed to spend more money than students from a rural background so the former received more money from the government. Girls received a lower amount because they were expected to make and repair their own clothes. Processing the study grants was

¹⁵ See Zouridis (2000).

¹⁶ See also Zouridis (2000) and Bovens and Zouridis (2002).

¹⁷ See Lipsky (1980).

not supported by technology apart from the typing machines used for administering the decisions.¹⁸ The decisions were signed in person by the street-level bureaucrats.

From the early twentieth century until the 1960s the configuration of legislation, organization, and technology was stable. The first step towards a new configuration was set with the mechanization of the administration.¹⁹ The first computers only supported the street-level bureaucrats on administrative matters such as printing and registering information. Gradually the computers crept into the core work of the street-level bureaucrats. Bureaucrats had to motivate the decisions they made which enabled the development of text blocks with standard motivations that allowed the street-level bureaucrats to more efficiently draw up decisions. Because of the growth of the student numbers, the organization was forced to better formalize the decision rules hence the legislation became more detailed. Gradually, a new configuration evolved in which the organization changed from a street-level bureaucracy to a screen-level bureaucracy. While the computers still supported the street-level bureaucrats, their work transformed from making decisions to putting the information in the computer system. Until 1986 the legislation still allowed street-level discretion, but the leeway for making decisions gradually faded because of ever more detailed rules.

In 1986 a new legislation took effect that marked the step towards a screen-level bureaucracy.²⁰ Although this was seen as a major step the configuration had already gradually changed because of the dynamics generated by the computers. The automation of the motivation of decisions required standard text blocks that in turn triggered the formalization of the considerations leading to a decision. Consequently, the formalization of the considerations enabled the development of an algorithm for decision-making. Creating a legal basis for the algorithm meant further elaboration of the legislation on study grants. The introduction of detailed legislation in 1986 marked a final step in this gradual process. The legislation introduced a basic study grant for every student who passed the age of 18 years. While the amount of the basic grant only depended on whether students lived with their parents or by themselves, the surplus could easily be calculated. The decision on the grant as well as the amount of the grant were now fully automated and the only work left for the officials was putting the data of the applications in the computer. The work of these screen-level bureaucrats mainly consisted of entering forms filled in by students on the computer. The computer decided and printed the decision by itself. The configuration shift that had taken place continued throughout the 1980s.

In the 1990s the organization again encountered a major set of changes.²¹ The computer system that decided on study grants gradually evolved from a series of mass production computer systems to a set of information systems connected with each other. Instead of having the screen-level bureaucrats entering the information,

¹⁸ See Zouridis (2000).

¹⁹ See Zouridis (2000).

²⁰ Wet op de studiefinanciering (1986).

²¹ See Zouridis (2000).

the same information could, for example, be extracted from the municipal register that contains the home address of the students. So why wait for students to apply for a study grant if both the registration at a university is known by the ministry as well as the moment the student turns 18 years old hence eligible for a study grant? The organization that took care of study grants even changed its name into Information Management Group. Gradually the screen-level bureaucrats also disappeared from the organization. They were replaced by system-level bureaucrats.²² These officials did not deal with individual cases anymore but only focused on the production process. Woodward's mass production was replaced by continuous process production. Like the staff in an oil refinery does not actually refine oil but only design and maintain the process of refining oil, the system-level bureaucrats were now only focusing on refining the information on students eligible for study grants. The algorithm on study grants had become an application linked with information systems in different government organizations. The system-level bureaucrats were busy managing the links between these systems as well as the maintenance of the algorithm. A new configuration evolved that required thorough harmonization of the legislation beyond the mere law that regulated study grants. Concepts like income and address had to be harmonized in many different laws in order to accommodate the system-level bureaucracy and its technology. The end of the 1990s marked the second configuration shift that turned the screen-level bureaucracy into a system-level bureaucracy.

2.2 *Traffic Violations*

Even though the configuration shifts do not overlap in time the same shifts have been observed in a completely different policy field. Traffic regulations enforcement in the Netherlands was traditionally regulated by criminal law and criminal procedure.²³ Street-level bureaucrats populated the entire process of deciding on the violation and the traffic fines (the police), the possible prosecution of these violations (the public prosecutor's office), the judicial decision on the violations (the court), and the collection of the penalties (the judicial agency that collected the money of these fines). Each case was assessed individually under criminal law which meant that all elements of the definition of the crime had to be established first. In case of speeding, the fact of speeding had to be determined but also the police had to determine who was driving the car because this was the person who committed the violation. For decades computers did not play any role in these processes apart from typing machines used for production of decisions. The street-level bureaucrats took care of all the operations manually. So, in case of a speeding violation, a police officer who detected a violation of the speed limit could decide to

²² See Bovens and Zouridis (2002).

²³ See Zouridis (2000).

write a speeding ticket. Criminal procedure legislation considers such a speeding ticket as a so-called transaction. If the driver pays the fine, he or she averts criminal proceedings. If he or she did not pay, the case was transferred to the public prosecutor's office. Since the workload usually was dramatically higher than the capacity of the public prosecutor's office many traffic violations were just dismissed by the public prosecutor or simply not forwarded by the police. If the public prosecutor decided to proceed with the case, he proposed a higher fine as a transaction. Again, the driver could choose to pay the fine or just wait. If he waited the case could be put forward to the subdistrict court. Many cases were just dismissed because of workload considerations. If the subdistrict court imposed a sanction (usually a fine) this was collected by the police or a collection agency. The story may become boring but if the convicted person did not pay the workload again superseded the capacity of the organization so in this stage many cases were just thrown away. Only in rare instances the police actually visited these convicted persons to collect the fine.

At the end of the 1970s computers were introduced to support these processes. The first steps towards computerization were to mechanize the administration and computerize the records. The computerization followed the administrative process within the separate organizations and the computers only supported the administration. The core task of deciding on traffic violations as well as prosecuting or convicting a violation remained untouched by the computer. Since criminal procedure requires an individual assessment of the circumstances of the case, computers only impacted the administrative support staff. While each organization increasingly used computers to store information on violations and logistically guided the cases through the administrative processes, the decision-making processes were not affected by computerization until the end of the 1980s.

In 1990 the entire configuration shifted from a system of unit production to a system of mass production. In the legislation, small traffic offences were transferred from criminal law to administrative law. While criminal law requires the police and the public prosecutor to prove the offence as well as to prove that the offender committed the offence the new system of administrative law introduced a system of liability of licence plate holders. Clearly, the violation still had to be proved but excuses of car owners such as that it wasn't them who committed the offence or any other excuse were now almost completely useless because the license plate holder was legally liable. The law removed the public prosecutor from the initial process and enabled large-scale processing of these cases. Cameras now increasingly registered the traffic violations, and the computer read the licence plates in order to produce administrative decisions printed by the same computer system. At first, administrative officials still had to input the license plates so at the start of the 1990s the system of street-level bureaucracies transitioned into a screen-level bureaucracy. A new national agency was set up for the collection of administrative fines. Citizens who disagreed with the administrative decision could still appeal to the public prosecutor and the court as a final resort. Since even the court was bound by administrative law which does not require an individual assessment because of the liability of car owners, appeal was usually ineffective. The only defence left for individual citizens is the improbable but not impossible option of errors regarding the license

Table 1 Characteristics of the configurations and the configuration shifts in the realms of social welfare legislation and criminal law

Configuration	Legislation	Organization	Technology
Street-level-centered configuration	Assessment-inducing legislation	Street-level bureaucracy (unit based production)	Support systems (data bases, text editing)
Production-centered configuration	Parametric legislation	Screen-level bureaucracy (mass production)	Production systems (case-handling systems)
Information-centered configuration	Harmonizing legislation	System-level bureaucracy (continuous production)	Connected networks of information/applications

plate. For instance, the licence plate can be misread by the computer, or a change of owner has not been registered yet by the government agency that issues the license plates.

The configuration shift immediately proved to be a great success, at least from the perspective of the government. While before the configuration shift approximately half of the traffic fines were paid, now both the number of fines as well as the collection rate multiplied.²⁴ The Ministry of Justice soon found out that also for other types of violations collecting and executing court decisions would improve the level of efficiency. The national agency for the collection of fines and the execution of court decisions (Centraal Justitieel Incassobureau) grew rapidly in the 1990s and it developed new computerized systems for new tasks. The agency now became a full-fledged screen-level bureaucracy. All the computer systems followed the logic of the production process (the input of data on a violation or a court decision, the production of the payment orders and the steps of raising the amount to be paid in case of non-payment with imprisonment as the ultimate step). Gradually the agency found out that the same citizens were sometimes approached several times a week by the agency because these citizens were at the same time liable for a traffic fine and a court decision. The agency was unaware of the concurrence because of the design of the computer systems as mass production systems for individual decisions. It increasingly felt the need to shift from a screen-level bureaucracy to a system-level bureaucracy in which the different computer systems are connected so information can be exchanged across computer systems.

These two cases differ with regard to the timing of the configuration shifts as well as regarding the completely different nature of their tasks and the legal context in which the agencies operate. While study grants are typical examples of welfare benefits and legislation on social security and social welfare the traffic fines belonged to the realm of criminal law and criminal procedure. Since these cases were selected because of the variation of policy fields in order to detect the impact of ICTs beyond the particular policy context the resemblances are striking. The ICTs triggered similar configuration shifts which are summarized in Table 1.

Since AI has only recently been introduced to support the operations of large-scale government bureaucracies it is too early to predict AI's impact on the

²⁴ See Zouridis (2000) for the exact numbers of fines.

configurations. The next section therefore explores the possible impact of AI given the characteristics of current AI technology.²⁵

3 AI and the Contours of a New Configuration

How will AI change the configuration of legislation, basic organization principles, and technology in the realm of large-scale bureaucracies in government? Although there are some ideas and some research has been published, the future of AI's impact is speculative.²⁶ In these large-scale bureaucracies AI is still sparsely used. Moreover, since the development of AI is at its early stages, it is hard to tell how it will evolve both as a core technology and as a set of applications to be deployed by large-scale bureaucracies. But we do not necessarily have to resort to irresponsible speculation and science fiction because the current AI does already provide a glimpse into the key features of this kind of technology. Whatever its future development it is already clear that AI should be regarded as a system technology that will fundamentally change and reshape society.²⁷ According to the Dutch Scientific Council on Government Policy²⁸ AI is a system technology because it is, or will be, embedded in all other technologies and applications ranging from cars to web shops, and from processes such as the diagnosis of illness to forecasting the weather. It may thus be expected that it will also change the basic processes and routines of large-scale bureaucracies that implement the law. As well as with other technologies we can take the basic characteristics of AI as a starting point to assess the configuration shift and the future configuration. We therefore first have to establish the basic characteristics of AI. In order to establish these basic characteristics it does not suffice to merely look at applications such as self-driving cars or medical appliances. Thus far artificial intelligence is a catch-all term for a variety of technologies such as self-learning algorithms, micro-targeting, and computer systems that mimic human intelligence.²⁹ In some ways, AI builds on the technologies that produced earlier configuration shifts such as the automated or algorithmic decision making that triggered the shift from street-level to screen-level bureaucracy or the networked information systems that triggered a further shift to system-level bureaucracy. Compared with networked automated decision-making, AI does add some new features although the intelligence of AI should not be overestimated.³⁰ AI is fundamentally different from earlier technologies because of three reasons. Compared with earlier

²⁵ See also Yeung (2022).

²⁶ See for ideas and research Bullock (2019), Flügge et al. (2021), Vredenburg (2023), Aneesh (2009), De Boer and Raaphorst (2023), Engin and Treleaven (2019), Young et al. (2019), Lorenz et al. (2021), Plantinga (2022) and Peeters and Widlak (2023).

²⁷ See WRR (2021).

²⁸ See WRR (2021).

²⁹ See WRR (2021).

³⁰ See Pearl and Mackenzie (2018).

algorithms, the AI algorithms are dynamic because of their capacity to learn due to machine learning techniques. Algorithms were already included in the technologies that produced the configuration shift from the human-centered to the production-centered configuration. Capturing the underlying legislative framework in a computer system required that the decision-making as laid down in legislation was translated into computer code so the computer could simply apply legislation by following the same decision rules. AI does not necessarily require reprogramming the software, but it is able to change the algorithm by itself. While previous technologies could not change the decision rules without human intervention (rewriting the software) the current AI technologies can learn by themselves and change the algorithm without human intervention. Some software for automated drivers assistance systems (ADAS), for example, learn from the responses of drivers to its algorithms and then changes the algorithm by itself.³¹ As the ADAS example shows there is a data driven feedback loop embedded in AI. AI learning takes place by analyzing the data fed to the computer system. A decision making algorithm can thus change because of the patterns analyzed using data from either the system itself or other information systems. The data-driven feedback loop based on pattern recognition characterizes AI and distinguishes it from earlier technologies. If tax authorities use AI to impose taxes on small businesses, AI can detect the fingerprints of fraudulent business owners and change the algorithm in such a way that the probability of a visit by a tax inspector for these business owners increases. The entire learning cycle that previously took quite some effort by both computers for storing and sharing data and humans who detected patterns and changed the algorithm is now embedded in the technology itself. A third reason to consider AI as something fundamentally different from earlier technologies is the nature of the man-machine interaction. As Pearle and Mackenzie argue AI is not as intelligent as even a three-year old human.³² It cannot, for example, discern a bicycle on a picture by itself. If humans train the system to discern a bicycle by indicating on thousands of pictures whether a picture does or does not contain a bicycle the computer is able to learn to detect a bicycle by itself if confronted with new pictures. Even in case of unsupervised learning techniques AI is not capable to decide on whether the learning that took place is actually correct or flawed without human judgment. AI cannot become intelligent without human-computer interaction because it needs human judgment, the outcome of previous cases, or a mechanism to verify whether the outcome matches reality (usually but not necessarily humans). So if the algorithm directs the tax inspector to suspicious small business owners the feedback from the inspector (was this business owner indeed fraudulent?) enables the algorithm to further improve itself by detecting the variables that correlate with the fraudulent small business owners discovered by the tax inspectors. This particular men-machine interaction is at least characteristic for current AI technologies. It can be doubted

³¹ See for example Dutch Safety Board (2019).

³² See Pearl and Mackenzie (2018).

whether future AI still requires human judgment, but current AI is still in its infancy years.

Because of these reasons we can now take the next step, which is to imagine its use in large-scale government bureaucracies that apply legislation. What will be the impact of AI on the configuration of organization, legislation, and technology in large-scale bureaucracies? The immediate result of AI as data-driven self-learning computer systems enriched by human judgement is an ever-more accurate decision-making process that at the same time will become ever-more untraceable and maybe even completely unpredictable. The continuous feedback loops and men-machine interactions feed the algorithm in such a way that the algorithm can change itself. The exact decision rules, data, and human judgment used for an individual decision will quickly become obscure or at least very difficult to reconstruct. The immediate decision rules can always be logged hence reconstructed but the combination of data, ever-changing decision rules, and the interaction between human judgment and the algorithm may eventually even become so difficult that any attempt to reconstruct it is fruitless. And even if the mix of drivers can be reconstructed it will be difficult if not impossible to assess the correctness by a human official (because the outcome cannot be determined) so probably the only option left is to rely on the outcome of AI. An experiment with police officers and AI shows that the street-level bureaucrats rely on AI if it aligns with their professional judgment.³³

The characteristics and their immediate impact on decision-making by large-scale government bureaucracies provide a glimpse of the configuration shift that may be expected as a result of AI. As argued above the configuration encompasses the core technology of the organization as well as the legislation and the key characteristics of the organization. So let's now imagine the technology, organization, and legislation of a large scale bureaucracy for welfare benefits or traffic fines that uses AI. The core technology is artificial intelligence or machine learning. Self-learning algorithms decide on welfare benefits by gathering the data necessary for the decision-making process from any database connected with the algorithm. Citizens do not need to apply for a welfare benefit because AI continuously scans the data on the conditions for a welfare benefit. The organizations that impose and collect traffic fines more or less operate the same way in an AI context. The algorithm continuously scans the data available from computer systems or even the computers in cars with which it is connected to check for the circumstances that trigger a traffic fine. Because of the self-learning nature of these algorithms, they also continuously learn from the feedback fed either by data or by human judgment. So if certain persons systematically do not pay for their traffic fines or if the traffic fine was imposed on the wrong license plate holder because of a time gap after selling the car the algorithm will learn from it and change the decision-making process for these categories. If specific persons who receive a welfare benefit or study grant commit fraud the algorithm is fed with the fraudulent outcome. Eventually the algorithm will recognize the characteristics of these individuals and learn by itself.

³³ See Selten et al. (2023).

Table 2 The future AI-driven configuration of legislation, organization, and technology

	Legislation	Organization	Technology
Continuous iterative pattern-centered configuration	Meta legislation	Learning-loop bureaucracy	AI/machine learning

Apart from maintaining the algorithms or the computer systems that run these algorithms the organization consists of bureaucrats who assess outcomes and, if necessary, train or re-train the algorithm. To assess the outcomes, the latter bureaucrats will have to connect with individual cases. The organizing principle for these large-scale bureaucracies is the continuous learning loop that connects the human organization parts with the technological parts.

Now what will be the implications for legislation? While the organization and the algorithms continuously learn either because of the data analysis and pattern recognition or because of the training by bureaucrats it may be expected that legislation also changes. At least two types of legislation will logically evolve. On an operational level the legislation copies the algorithm so legislation will become as self-learning or even volatile as the algorithm itself. Because of the legal uncertainty that will arise and the need for some legal stability and coherence, a new type of legislation will evolve on a strategic level. Legislation on any substance such as study grants or traffic fines will become meta-legislation that does not regulate the content of the algorithm but the learning process of the algorithms. Meta-legislation may, for example, regulate the transparency of the learning loop as well as checks and balance in the human-machine learning that takes place in the organization. The new configuration that evolves can be labeled as a continuous iterative pattern-centered configuration since all components (legislation, organization, and technology) are focused on patterns in the data on individual cases. Table 2 summarizes and captures the new configuration of learning-loop bureaucracies as triggered by AI.

4 Constitutionalizing Learning-Loop Bureaucracies

For more than an age, constitutionalizing the bureaucracy has been a major concern in political science, public administration, and constitutional law. The power of large-scale bureaucracies is embedded in the information asymmetry between democratic actors and the bureaucratic agents. Ever since the early twentieth century scholars have come up with theories on how to constrain the bureaucracy.³⁴ For many decades these scholars focused on the street-level bureaucrats who decided on individual cases. At least three strategies have been used to constitutionalize the work of street-level bureaucrats:³⁵

³⁴ See Weber (1922), Hayek (1944, 1960) and Waldo (1948).

³⁵ See Cane (2016) and Zouridis (2021).

1. Detailed legislation: with ever-more detailed legislation the legislators have tried to constrain street-level bureaucrats. By replacing open norms with more precise decision-making parameters and by simply adding rules that should be taken into account when deciding on individual cases the work of street-level bureaucrats has been enormously constricted.
2. Procedural legislation: if legislators are not able to precisely set the parameters for street-level decision-making, they may resort to procedural norms to regulate the decision-making process. There are many examples of procedural norms such as the obligation to motivate decisions, to first hear the citizens involved, or to first consult an advisory body.
3. Checks & balances: several checks and balances have been developed as an additional means to constrain street-level bureaucrats. While common law systems already had the judicial check by the courts, both the US and continental European law further developed systems of administrative procedure and judicial review even to the extent that some commentators argue that the current set of administrative procedures should be regarded as unconstitutional³⁶ or as counterproductive.³⁷

The rise of screen-level and system-level bureaucracies has attracted some attention of public administration and legal scholars on how to constitutionalize these new types of bureaucracies.³⁸ Only recently the explosion of algorithms in society also triggered more attention to the use of algorithms in government bureaucracies because of scandals³⁹ or because of the racial and other biases embedded in the algorithms.⁴⁰ So how can the learning-loop bureaucracy be effectively constitutionalized? Which of the strategies mentioned above can be used to constitutionalize the new AI-configuration?

First, we have to establish what's at stake. What is the rationale for constitutionalizing? In the old street-level bureaucracy the power of street-level bureaucrats was at stake. Their discretionary power in the cases described above meant that the street-level bureaucrats could decide to give or refuse a welfare benefit or to impose a fine following a traffic violation. Detailed legislation, rules, procedures, and checks and balances gradually constrained the discretionary power of street-level bureaucrats. What should be constrained in the era of AI? Each configuration shift also implied a shift of decision-making power. The shift from street-level to screen-level bureaucracy shifted discretion and decision-making power from the street-level bureaucrats to the programmers who developed the algorithms in the computer system by writing the computer code. The shift from screen-level to system-level bureaucracy again shifted discretion and decision-making power to the software designers who designed the networked information systems because their choices

³⁶ See Hamburger (2014).

³⁷ See Kagan (2001).

³⁸ See Bovens and Zouridis (2002) and Krafft et al. (2020).

³⁹ E.g. in the Netherlands, see Peeters and Widlak (2023).

⁴⁰ See O'Neil (2016).

were embedded in the computer networks that produced the decisions. It may be expected that the shift from system-level to pattern-level again causes a shift of discretionary power and decision-making. First of all, self-learning algorithms absorb discretionary power and decision-making powers. While the computer programmers and software designers could fixate decision-making in algorithms, AI can only partially be influenced by the humans who program and operate the algorithms. So, the self-learning algorithms themselves need to be constrained first. In this respect, meta-legislation also entails rules on the learning process of the algorithms hence regulating the machine learning itself. All three strategies mentioned above can be used to regulate machine learning. The first strategy, clear substantive standards, regulates what learning processes are allowed (e.g. ‘doppelganger’ analysis) and what learning processes are prohibited (e.g. racial profiling)? Or: to what extent can ‘doppelganger’ or other techniques be used for individual targeting?⁴¹ What types of data are allowed to be used by AI to construct patterns?

The second strategy focuses on the procedures that have to be followed in the case of machine learning. While transparency of the algorithm and the learning process may be difficult to regulate in a volatile AI environment legislation could include logging obligations to detect when and what data were used, and by which decision rules, to produce an individual decision. Finally, the strategy of checks and balances can also be used, both with human and technological checks and balances. Given the fundamental unpredictability of AI generated content an AI generated second opinion may even be considered.

It is plausible that entering the era of AI will produce a configuration shift. This paper explored the nature of the configuration shift that may be expected as some strategies to constitutionalize the new configuration. It appears that AI may both cause a shift of decision-making and discretionary power as well as provide the tools to constitutionalize it. AI may not only lead to the most perfect bureaucracy ever dreamt of by Weber, Hayek, and Waldo it may even enable us to better constrain these bureaucracies than they could imagine.

References

- Aneesh A (2009) Global labor: algocratic modes of organization. *Social Theory* 27(4):347–370
- Bovens M, Zouridis S (2002) From street-level to system-level bureaucracies: how information and communication technology is transforming administrative discretion and constitutional control. *Public Adm Rev* 62(2):174–184. <https://doi.org/10.1111/0033-3352.00168>
- Bullock JB (2019) Artificial intelligence, discretion, and bureaucracy. *Am Rev Public Adm* 49(7):751–761
- Cane P (2016) *Controlling administrative power. An historical comparison*. Cambridge University Press
- Chandler AD Jr (1962) *Strategy and structure: chapters in the history of the industrial enterprise*. MIT Press

⁴¹ See for example Stephens-Davidowitz (2017) on the different types of data analysis.

- de Boer N, Raaphorst N (2023) Automation and discretion: explaining the effect of automation on how street-level bureaucrats enforce. *Public Manag Rev* 25(1):42–62. <https://doi.org/10.1080/014719037.2021.1937684>
- Dunsire A (1978) The execution process: implementation in a bureaucracy: v. 1. Martin Robertson
- Dutch Safety Board (2019) Who is in control? Road safety and automation in road traffic. Dutch Safety Board. <https://onderzoeksraad.nl/en/onderzoek/who-is-in-control-road-safety-and-automation-in-road-traffic/>
- Engin Z, Treleaven P (2019) Algorithmic government: automating public services and supporting civil servants in using data science technologies. *Comput J* 62(3):448–460
- Flügge AA, Hildebrandt T, Møller NH (2021) Street-level algorithms and AI in bureaucratic decision-making: a caseworker perspective. In: Nichols J (ed) *Proceedings of the ACM on Human-Computer Interaction*, vol 5 (art. 40). Association for Computing Machinery, pp 1–23. <https://doi.org/10.1145/3449114>
- Hamburger P (2014) *Is administrative law unlawful?* The University of Chicago Press
- Hayek FA (1944) *The road to serfdom*. Routledge
- Hayek FA (1960) *The constitution of liberty*. Routledge and Kegan Paul
- Kagan RA (2001) *Adversarial legalism: the American way of law*. Harvard University Press
- Kelly K (2010) *What technology wants*. Viking
- Krafft TD, Zweig KA, König PD (2020) How to regulate algorithmic decision-making: a framework of regulatory requirements for different applications. *Regul Gov* 16(1):119–136. <https://doi.org/10.1111/rego.12369>
- Lipsky M (1980) *Street-level bureaucracy: dilemmas of the individual in public services*. Russel Sage Foundation
- Lorenz L, Meijer AJ, Schuppan T (2021) The Algocracy as a new ideal type for government organizations: predictive policing in Berlin as an empirical case. *Inf Polity* 26(4):71–86
- Maynard-Moody S, Musheno M (2003) *Cops, teachers, counselors: stories from the front lines of public service*. The University of Michigan Press
- Mintzberg H, Ahlstrand B, Lampel J (1998) *Strategy Safari. A guided tour through the wilds of strategic management*. Prentice Hall
- O’Neil C (2016) *Weapons of math destruction*. Crown Books
- Pearl J, Mackenzie D (2018) *The book of why: the new science of cause and effect*. Basic Books
- Peeters R, Widlak AC (2023) Administrative exclusion in the infrastructure-level bureaucracy: the case of the Dutch daycare benefit scandal. *Public Adm Rev* 83(4):863–877
- Plantinga P (2022) Digital discretion and public administration in Africa: implications for the use of artificial intelligence. *Inf Dev* 40(2):332–352
- Prottas JM (1979) *People processing: the street-level bureaucrat in public service bureaucracies*. Lexington Books
- Selten F, Robeer M, Grimmelikhuijsen S (2023) ‘Just like I thought’: street-level bureaucrats trust AI recommendations if they confirm their professional judgment. *Public Adm Rev* 83(2):263–278
- Stephens-Davidowitz S (2017) *Everybody lies. Big data, new data, and what the internet can tell us about who we really are*. HarperCollins Publishers
- Thompson JD (1967) *Organizations in action. Social science bases of administrative theory*. McGraw-Hill
- van der Pot JHJ (1994) *Steward or Sorcerer’s Apprentice? The evaluation of technical progress: a systematic overview of theories and opinions*. Eburon
- Vredenburg K (2023) AI and bureaucratic discretion. *Inquiry* 66:1–30. <https://doi.org/10.1080/0020174X.2023.2261468>
- Waldo D (1948) *The administrative state: a study of the political theory of American Public Administration*. Ronald Press Company
- Weber M (1922) *Wirtschaft und Gesellschaft. Grundriss der verstehenden Soziologie*. Paul Siebeck
- Wilson JQ (1989) *Bureaucracy. What government agencies do and why they do it*. Basic Books
- Winner L (1980) Do artifacts have politics? *Daedalus* 109(1):121–136

- Woodward J (1958) Management and technology. Her Majesty's Stationery Office
- WRR (Scientific Council for Government Policy) (2021) Opgave AI. De nieuwe systeemtechnologie (Mission AI. The New System Technology.). WRR
- Yeung K (2022) The new public analytics as an emerging paradigm in public sector administration. *Tilburg Law Rev* 27(2):1–32. <https://doi.org/10.5334/tilr.303>
- Young MM, Bullock JB, Leczy JD (2019) Artificial discretion as a tool of governance: a framework for understanding the impact of artificial intelligence on public administration. *Perspect Public Manag Gov* 2(4):301–313
- Zouridis S (2000) Digitale disciplineren. Over ICT, organisatie, wetgeving en het automatiseren van beschikkingen (Digital discipline. On ICT, organization, legislation and automated government decisions) [Doctoral Dissertation, Tilburg University]. Eburon
- Zouridis S (2021) De resonante rechtsstaat (A resonant rule of law). BoomBestuurskunde

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

