Chapter 9

The Rule of Law for Artificial Intelligence in Public Administration: A System Safety Perspective*

Sem Nouws & Roel Dobbe

Abstract This chapter proposes an analytical lens to comprehensively address the role of Artificial Intelligence (AI) applications in mediating arbitrary exercise of power in public administration and the citizen harms that result from such conduct. It provides a timely and urgent account to fill gaps in conventional Rule of Law thought. AI systems are socio-technical by nature and, therefore, differ from the text-driven social constructs that the legal professions dealing with Rule of Law issues concentrate on. Put to work in public administration contexts with consequential decision-making, technical artefacts can contribute to a variety of hazardous situations that provide opportunities for arbitrary conduct. A comprehensive lens to understand and address the role of technology in Rule of Law violations has largely been missing in literature. We propose to combine a socio-legal perspective on the Rule of Law with central insights from system safety – a safety engineering tradition with a strong scientific as well as real-world practice – that considers safety from a technological, systemic, and institutional perspective. The combination results in a lexicon and analytical approach that enables public organisations to identify possibilities for arbitrary conduct in public AI systems. Following on the analysis, interventions can be designed to prevent, mitigate, or correct system hazards and, thereby, protect citizens against arbitrary exercise of power.

Keywords Rule of Law; public algorithmic systems; arbitrary use of power; artificial intelligence; system safety; citizen harms

---

* To appear in T.C.M. Asser and Springer Information Technology and Law Series, special issue on 'Digital Governance: Confronting the Challenges Posed by Artificial Intelligence'

Sem Nouws
Delft University of Technology
Jaffalaan 5 2628 CN, Delft, The Netherlands
e-mail: S.J.J.Nouws@tudelft.nl

Roel Dobbe
Delft University of Technology
Jaffalaan 5 2628 CN, Delft, The Netherlands
e-mail: R.I.J.Dobbe@tudelft.nl

9.1 Introduction

Governments often fall short in adequately protecting citizens from harm inflicted through their dependency on algorithmic applications for the provision of public services. In the Netherlands, childcare benefit recipients were falsely accused of fraud based on a risk indication model.[4] Similarly, the Robodebt program in Australia falsely assigned debts to citizens based on automatically calculated overpayments.[5] False accusations and the subsequently imposed harsh penalties have resulted in grave harm done to unguilty citizens, often ruining lives and amounting to significant human right violations. Both cases are exemplary for the increased use of algorithmic and data-driven applications in the public sector. Especially the introduction of Artificial Intelligence (AI) applications such as machine learning illustrates a new step in automating processes in public administration by facilitating fraud prediction, risk assessment, and allocation of public resources. The exemplary cases of the childcare benefit scandal and the Robodebt scheme show that public organisations seem to fall short in preventing, mitigating, or correcting citizen harms inflicted by the use of algorithmic applications.

This chapter focuses on the use of AI in a public administration context. First, AI applications are rule- and/or data-driven software-based technologies that predict or generate output. These applications are used to automate or augment processes in public administration, which we refer to as AI practices. Second, it is important to stress that AI applications are one of several different components in a system. They are part of a larger bureaucratic environment that aims to provide a public service or execute public policy.[6] This means that AI applications are operated by human agents – e.g., public servants – that interact with or otherwise monitor the technology.[7] Moreover, the AI applications are situated in an institutional context – i.e., a set of formal and informal social rules that structure human behaviour and the technical specification of AI applications.[8] In the rest of this chapter, we will refer to this broader and systemic understanding as *public AI systems*.

Where public organisations use these applications to increase efficiency or address complex policy issues, their institutional and administrative practices are also challenged.[9] For example,

[4] Peeters and Widlak 2023.
[5] Braithwaite 2020.
[6] Mulligan and Bamberger 2019.
[7] Fountain 2001.
[8] Orlikowski 1992.
[9] Veale and Brass 2019.

practices are changed by creating opaqueness or changing power dynamics. In the already complex web of public services, AI applications can make the impact to households harder to predict and understand, e.g. by automating the determination of repayments. Changing power dynamics can be observed in the shift of discretionary power from the frontline worker to the developer of AI applications.[10] Moreover, these applications intervene in the fundamental relationship between citizen and government by formalising the structuration of citizen behaviour.[11] Consequently, the power imbalance between government and citizens increases;[12] creating possibilities for arbitrary use of power.[13] In cases of automated decision-making, public organisations often offload responsibility for proving eligibility and receiving benefits to citizens.[14] Besides, the necessary knowledge needed to navigate complex benefit systems, often consisting of a multitude of policies and services with various dependencies, can amount to an impossible barrier for citizens.

Disruption of institutional practices following from the use of AI applications also interferes with Rule of Law practices – i.e., efforts to reduce arbitrary use of power. Several authors have argued that AI and other forms of automation are a challenge or threat to the Rule of Law.[15] Others discuss the role of technology or AI applications in arbitrary use of power.[16] Current approaches to the Rule of Law seem to fall short in addressing issues arising from public AI systems as they do not integrate institutional and technical expertise.[17] However, apart from calls for ideas like 'legal protection by design',[18] it is still unclear how such an integration can be achieved in practical terms.

In this chapter, we develop an analytical lens that provides a way to detect and correct arbitrary conduct in contexts where AI applications are deployed to administer public services. We draw a relationship between protecting citizens against the system hazards in public AI systems and reducing arbitrary conduct. For this aim, we employ the socio-legal perspective on

[10] Bovens and Zouridis 2002; Alkhatib and Bernstein, 2019.
[11] Janssen and Kuk 2016.
[12] Yeung 2018.
[13] Brownsword 2016; Zalnieriute et al. 2019.
[14] Widlak and Peeters 2020.
[15] Bayamlıoğlu and Leenes 2018; Hildebrandt 2018; Greenstein 2022.
[16] Brownsword 2016; Zalnieriute et al. 2019.
[17] Zalnieriute et al. 2019.
[18] Hildebrandt 2011.

the Rule of Law suggested by Krygier and Selznick,[19] situating public service decision-making and its risk of arbitrary conduct in a social context in combination with a system safety approach. The latter leans on decades of insights from safety engineering and (socio-technical) systems theory about dealing with system hazards in software-based automation in complex processes.[20] The resulting conceptualisation opens new avenues to interpret and enrich the socio-legal perspective to account for the causal ways in which AI applications restructure and impact (arbitrary) conduct, as well as for the role of organisational and systemic factors contributing to such causal outcomes.

This chapter starts with discussing the need for a socio-technical perspective in the Rule of Law. In Sect. 9.2, we unravel the role of AI systems in arbitrary use of power by linking these systems to manifestations of arbitrary conduct. Sect. 9.3 examines how Rule of Law perspectives understand the role of technical artefacts in arbitrary conduct and how that understanding disregards the structuring nature of technology. Thereafter, in Sect. 9.4, we show the merits of including a socio-technical perspective in the Rule of Law and argue for the suitability of system safety as a specific socio-technical perspective. The subsequent sections discuss three focus points on which the socio-legal perspective and system safety could be combined: a lexicon, an analytical approach, and safety-guided design. Sect. 9.5 presents a lexicon that rephrases dominant conceptualizations of arbitrary use of power through a system safety lens and adjacent systems engineering concepts. Applying the system safety lens to public sector decision-making and law execution involving AI systems, can then open up various avenues presented in Sect. 9.6 to analyse safety hazards – i.e., possibilities for arbitrary conduct – in such processes. From there we identify readily available methods to address such issues, as well as opportunities to extend system safety approaches to be of value in addressing Rule of Law issues related to public AI systems. In Sect 9.7 we show how known measures from the Rule of Law can be appropriated to system safety control actions. These action either prevent arbitrary conduct from occurring in the process or instigate mitigating measures in the operation of the process or the surrounding organizational governance.

[19] Selznick 1999; Krygier 2009.
[20] Leveson 2012.

9.2 Arbitrary conduct mediated by public AI systems

To address arbitrary use of power mediated by public AI systems, the role of such systems in arbitrary conduct needs to be clarified. However, arbitrary use of power as a concept itself is under-theorised.[21] Mak and Taekema, and Krygier have provided first attempts to categorise and define arbitrary conduct.[22] We base our discussion of arbitrary conduct in public AI systems on four – non-exclusive and non-exhaustive – manifestations of arbitrariness listed by these authors: (1) reasoning by public servants based on 'own will or pleasure'; (2) inability of citizens to engage in or contest decision-making; (3) unpredictability and incomprehensibility of conduct for those affected; and (4) unfair decision-making in concrete situations. This section exemplifies how these four manifestations may emerge in public AI systems.

9.2.1 Reasoning based on own will or pleasure

Conduct is arbitrary if it is based on the own will or pleasure of an individual exercising power. This means that a rational basis or good arguments for a decision is missing.[23] Such a rational basis is formed by formal rules and regulations, procedures, or mandates and steers or limits the power exercised by decision-makers.[24]

Public AI systems enable two types of actors to impose their own will or pleasure on others: the operator(s), that is, the public servant deploying the system or using its output, and the designer(s) of the system, that is, the actors responsible for specifying or developing the system. The system's operator can intentionally use the system to 'rationalise' their decision based on own will or pleasure, thereby possibly overriding other rational arguments. AI systems may also be used selectively to confirm one's own biases.[25] Another issue arises when operators blindly follow faulty AI output, not using their own critical reasoning to prevent an undesirable outcome. This phenomenon is sometimes referred to as automation bias, and hints at the risk of deskilling and loss of operational discretion.[26] However, the operational actors are not always to blame.[27]

[21] Krygier 2016.
[22] Krygier 2016; Mak and Taekema 2016.
[23] Mak and Taekema 2016.
[24] Krygier 2016.
[25] Young et al. 2021.
[26] Green and Chen 2019.
[27] Green 2022.

Often, errors are a function of the environment in which an operator is acting.[28] In other words, operators often lack the right information or resources to oversee or work with an AI system. In such cases, discretion of the system's operator at the operational level is effectively reduced or distorted by design choices made earlier in the system's life cycle.[29]

Similarly, AI system's designers can base their design choices on their own will or pleasure.[30] A special situation in this respect is the case that rules are not appropriately translated to the logical innerworkings of an AI system. This can be the result of, for example, vagueness or the urge to simplify inherent complex situations in formalistic models.[31] Thereby, system designers may impose a system's logic that does not align with laws, regulations, and policies related to the processes in which the AI system functions.

9.2.2 No space or means to engage or contest

Citizens should be able to engage in or contest decision-making through possibilities to question or voice arguments and complaints. An important feature of exercising power is that the interests of individuals affected are considered.[32] Therefore, those affected should also have the means and possibility to control and question those in power, and to be heard by them.[33]

Citizens are unable to engage in design processes of or contest AI systems when the responsibilities for such a system are poorly specified. AI applications are known for the responsibility gaps these create.[34] For example, AI systems are dependent on datasets and information architectures that might be situated in other organisational units.[35] In fact, AI systems often rely on vast and potentially global supply chains,[36] with inherent complexities that contribute to developers or users not experiencing or taking responsibility.[37] In the same vein, AI

[28] Leveson 2012.
[29] Cf. Peeters and Widlak 2018; Zouridis et al. 2020; Leveson 2012.
[30] König and Wenzelburger 2021.
[31] Dobbe et al. 2021; Alkhatib 2021.
[32] Mak and Taekema 2016.
[33] Krygier 2016.
[34] Santoni De Sio and Mecacci 2021.
[35] Sculley et al. 2015.
[36] Cobbe et al. 2023.
[37] Widder and Nafus 2023.

applications may be put in place to widen the *accountability gap* 'between those who develop and profit from AI—and those most likely to suffer the consequences of its negative effects.'[38]

As argued above, the use of AI systems in public administration is shifting power dynamics in public organisations, for example, by shifting discretion from bureaucrats to system designers,[39] or by strengthening the relative position of executive branches in governments.[40] Hence, assigned responsibilities may not reflect the real influence of actors on the AI system. A mounting problem in this context is the handover of development of AI systems – that intimately mediate public services – to external parties. Thereby, shifting both public accountability as well as autonomy over the quality of the public services from public to private actors.[41] Furthermore, normative design choices may be arbitrary if these are not based on adequate deliberation or (political) mandates.[42]

9.2.3 Unpredictable and incomprehensible

For conduct not to be arbitrary, citizens should be able to comprehend rules in order to comply with these rules.[43] This also means that rules and their enforcement should be predictable.[44] The requirements of predictability and comprehensibility also support citizens in challenging arbitrary conduct (see 1.2.2).

The opaqueness and complexity inherent to AI systems impedes comprehensibility and predictability of power exercised through public AI systems.[45] Citizens often lack the expertise needed to understand the working of AI systems.[46] Similarly, software-based automation systems can become so complex that operators also run into the limits of cognitive capacity to properly understand how the system functions and what behaviours might emerge under particular circumstances.[47] For example, semi-automated systems may look like the only effective and efficient way to compute and administer eligibility to and height of social welfare policies.

---

[38] Whittaker et al. 2018.
[39] Zouridis et al. 2020.
[40] Passchier 2020.
[41] Whittaker 2021.
[42] Hildebrandt 2011; Yeung 2014; Grimmelikhuijsen and Meijer 2022.
[43] Krygier 2016.
[44] Mak and Taekema 2016.
[45] Burrell 2016.
[46] De Bruijn et al. 2022.
[47] Leveson 2012.

Similarly, AI applications may also emerge to address the lack of predictability of complex social welfare systems. However, such semi-automated systems may quickly add their own complexity or their behaviour may turn unpredictable.

9.2.4 Unfair decision-making in concrete situations

Finally, Mak & Taekema stress the fact that unfair decisions can also be arbitrary. To prevent or reduce arbitrariness, decision-makers need to make contextual assessments of concrete situations.[48] This is related to considering the voices and interests of citizens in exercising power.[49]

AI systems can contribute to unfair decisions in concrete cases.[50] Discrimination and biases – prevalent in data-driven algorithmic systems such as AI –[51] may result in unfair decision-making. For example, when AI systems are used for allocation of benefits, biases may lead to allocation to citizens on basis of irrelevant characteristics. These biases may find their roots in historical conduct, as well as in the design choices made by developers, or in the ways in which the AI application is used.[52] Likewise, errors, flaws and the statistical or correlational nature of decision-making in an AI system can exclude citizens because of its disciplining nature – in which the consequences of such errors arbitrarily affect citizens.[53]

9.3 The lack of a socio-technical perspective in the Rule of Law

Sect. 9.2 shows that the role of AI applications in arbitrary conduct may not be the decisive factor, but cannot be ignored. When understanding the Rule of Law as governing people through law instead of by bureaucrats in order to reduce arbitrary use of power, the Rule of Law also applies to public AI systems. We discuss this role of the Rule of Law in this section. We show that conventional and current Rule of Law thinking lacks ways to bring into view and address the roles and implications of AI applications in arbitrary conduct.

The misconception of technologies' role in arbitrary conduct can be identified by considering the Rule of Law from a structuration lens. Structuration is about the duality or co-constitution of

[48] Mak and Taekema 2016.
[49] Krygier 2016.
[50] Barocas and Selbst 2016; Dobbe et al. 2018.
[51] Hildebrandt 2019.
[52] Dobbe et al. 2018.
[53] Peeters and Widlak 2018.

the two components institutions (structure) and human agents (agency).[54] These two components shape one another through interaction. Institutions are social rules that structure human behaviour.[55] In discussing institutions, legal literature mostly refers to legal institutions, such as rules, laws, rights, and procedures. This chapter broadens this conception and includes institutions such as culture, norms, and routines. By enacting institutions, human agents also change the form and function of institutions. Orlikowski has applied structuration to socio-technical systems.[56] Her model adds an extra component to the duality of institutions and human agents: technical artefacts. These artefacts are also in a co-constituting relationship with institutions and human agents.

9.3.1 Rule of Law in public AI systems

Unsurprisingly, various scholars have stressed the threats that public AI systems pose to the Rule of Law – which traditionally is considered the main mechanism to reduce arbitrary conduct.[57] Bayamlıoğlu and Leenes identify challenges to law as a normative, a causative as well as a moral enterprise and even caution 'that the 'rule of law' might be exchanged for the 'rule of technology' – accompanied by Kafkaesque, Huxleyan, and Orwellian discourses of dystopia.'[58] Furthermore, Hildebrandt addresses how machine learning applications challenge contestability, a fundamental element of Rule of Law.[59] Correspondingly, the use of AI systems in the public domain can undermine the Rule of Law.

On the other hand, the Rule of Law provides a framework to address arbitrary conduct mediated by AI systems. As mentioned, measures and institutions based on the Rule of law should be enacted to protect citizens against arbitrariness.[60] Several authors have argued that although public AI systems are of a different order compared to traditional laws and policy execution, they should fall under the Rule of Law regime. Brownsword stresses the 'continuing link between the regulators' normative intentions [represented in rule-based regulatory instruments] and the translation of these intentions into a technologically managed environment'

[54] Giddens 1976, Giddens 1979, Giddens 1984.
[55] Hodgson 2006
[56] The Structurational Model of Technology by Orlikowski 1992.
[57] See Krygier 2009.
[58] Bayamlıoğlu and Leenes 2018, p. 305.
[59] Hildebrandt 2016.
[60] Raz 1979; Krygier 2009.

–public AI systems can create such an environment.[61] Similarly, Hildebrandt argues that data-driven applications that execute laws count as mere administration (and not law), but that administration should also adhere to the Rule of Law.[62]

Consequently, AI applications used in the administration of law are subject to the Rule of Law, but their characteristics may simultaneously undermine Rule of Law mechanisms that are put in place to restrain the processes in which these technologies are applied. As such, the Rule of Law as a regime to reduce arbitrary conduct needs to be reassessed and adapted to ensure public AI systems operate under that regime and do not cause new issues to it. The need for broadening Rule of Law thinking also follows from the fact that it was not able to protect the citizens in the cases discussed at the start of this chapter. The toolboxes that perspectives on the Rule of Law provide did not enable governments to prevent, mitigate, or correct harms by arbitrary conduct in public AI systems.

9.3.2 Gaps in current Rule of Law perspectives

The Rule of Law is discussed from different perspectives in legal philosophy. First, we discuss the gaps in the *classic* perspectives in addressing public AI systems. Later, we discuss the socio-legal perspective.

The classic perspectives comprise the three angles – i.e., formal, substantive and procedural – from which scholars traditionally study the Rule of Law.[63] These three perspectives have different views on the nature of the Rule of law; each containing relevant and useful insights for governing public AI systems. The formal perspective provides (lists of) requirements for the form of rules for them to contribute to or comply with the Rule of Law.[64] The substantive perspective perceives the Rule of Law as guiding, demarcating, or constraining the content of rules. It emphasises the connection between rules and the moral and political rights that citizens have. The procedural perspective emphasises the Rule of Law as structuring argumentative practices that achieve objectivity. In other words, the Rule of Law should provide procedures to ensure that such an argumentative practice – e.g., in law-making or in court – runs properly.[65]

[61] Brownsword 2016, p. 102.
[62] Hildebrandt 2018.
[63] Cf. Waldron 2011.
[64] Raz 1979.
[65] Waldron 2011.

Although the three classic perspectives differ in their interpretation of the nature of the Rule of Law, they all focus on the form and function of legal institutions, and how these institutions structure human behaviour. Fig. 9.1 visualises this one-way interaction between the institutional component in social systems and human agents. The presupposition in these perspectives seems to be that human agents can engage in arbitrary conduct, but that this can be prevented, mitigated or corrected by establishing the right legal institutions.[66] The effects of human behaviour are thereby overlooked.

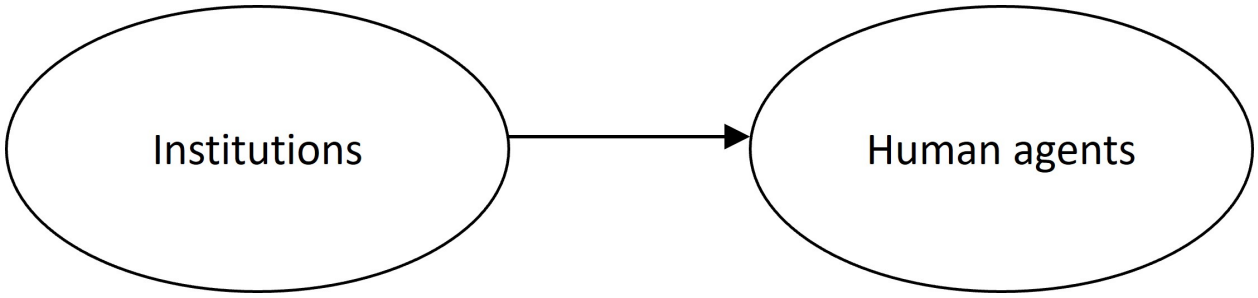

Figure 9.1 Structuration in 'classic' perspectives on the Rule of Law: institutions structure the behaviour of human agents[67]

The socio-legal perspective considers the same two components but acknowledges that structuration goes in both directions. This perspective developed amongst others by Selznick and Krygier uses insights from sociology to better understand how the Rule of Law can be achieved in practice by acknowledging that institutions are enacted by human agents.[68] It studies the effect that human behaviour has on the realisation of the Rule of Law. Fig. 9.2 shows this bidirectional interaction between the institutions and human agents.

Consequently, the perspective not only looks at legal institutions to sustain or materialise the Rule of Law, but has a broader perspective that includes political, administrative, and cultural aspects that organise and determine the functioning of the state. Therefore, a strong institutional context not only consists of legal institutions such as rules, but is backed by informal institutions such as culture, routines, and practices.[69] According to Krygier the 'ability to restrain the ways in

[66] Cf. Krygier 2009
[67] Figure by authors
[68] Selznick 2003; Krygier 2014.
[69] Nonet and Selznick 2001; Krygier 2009; Taekema 2021.

which power is exercised needs to be institutionalised.'[70] As such, while the socio-legal perspective takes the role of human behaviour into consideration, it still focuses on institutions as the main point of intervention for upholding the Rule of Law.

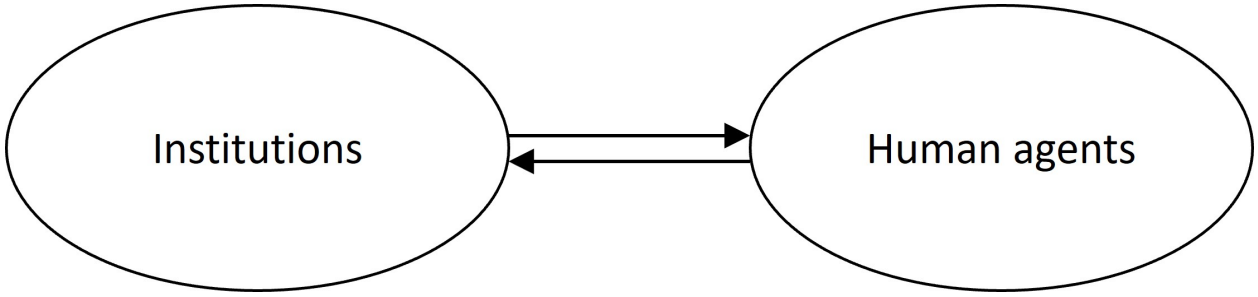

Figure 9.2 Structuration in the socio-legal perspective on the Rule of Law: human agents also structure the form and function of institutions[71]

We can hence conclude that both the classic as well as the socio-legal perspective on the Rule of Law do not explicitly consider technical artefacts, such as AI applications, as having a structuring function on human behaviour or institutions. Put differently, the perspectives consider institutions – in a narrow or broad sense – as the main point of intervention in establishing the Rule of Law. The importance of the right institutional context for public AI systems is also discussed outside legal literature.[72] Nevertheless, an institutional approach to the Rule of Law in a context mediated, informed or automated by AI applications will fall short in addressing the problems that public AI systems pose if it does not: (1) consider the ways in which AI practices conflict with Rule of Law principles, as well as (2) account for how these practices pose problems by structuring the administration of law and possibilities for arbitrary conduct. Put differently, public organisations need to have a granular understanding of how new technological applications, be they AI or other, structure their core processes and practices.

[70] Krygier 2009, p. 12.
[71] Figure by authors
[72] E.g., Dobbe 2022; Green 2022; Grimmelikhuijsen and Meijer 2022.

9.3.3 Public AI systems as socio-technical systems

Mostly, the role of technical artifacts is considered superficially or narrowly in the literature on the Rule of Law.[73] Consequently, the role of AI applications in arbitrary conduct is hard to understand from a purely legal philosophy standpoint. Existing socio-legal interpretations would benefit from a socio-technical structuration perspective, which assumes that an application of technology does not only comprise a technical artefact – i.e., the AI application –, but forms a system with institutions and human agents. System practices and outcomes hence emerge from the interaction between social and technical components when enacted by human agents.[74] Together, the technical, institutional, and agential components create possibilities for arbitrary conduct. Moreover, AI systems are embedded in a broader information architecture – i.e., they are connected to, interdependent with, and interrelated with other technical artefacts or systems.[75]

The three socio-technical components structure each other's form and function, as depicted in Fig. 9.3.[76] The interaction between institutions and technical artefacts is most apparent in the fact that public AI systems are constituted based on formal institutions such as laws, regulations, and policies. Moreover, the form and function of technical artefacts is also influenced by work instructions on how users should use the system. The other way around, the form and function of technology can, for example, institutionalise specific practices.[77] When human agents interact with technical artefacts their behaviour is disciplined by these artefacts. Likewise, human agents assign functions to artefacts by enacting public AI systems within the context of laws and policies by human agents (e.g., civil servant or front worker).[78] In other words, the AI application mediates the tasks or work of human agents and, thereby, can provide the agents with possibilities for arbitrary conduct when using or designing these systems.

Distinguishing the three components of public AI systems and examining their interactions enriches the understanding of the intricate role technical artefacts may play in arbitrary conduct.

---

[73] The few authors that write about AI and the Rule of Law do that from a specific focus or consider technology as an external or deterministic factor; e.g., Chiao 2023 has a focus on courts, Yeung 2018 as well as Cuéllar & Huq 2022 focus on regulation, Hildebrandt 2018 focuses on contestability.
[74] Orlikowski 1992.
[75] Nissenbaum 2019.
[76] Orlikowski 1992.
[77] Orlikowski 1992.
[78] Cf. Seaver 2017.

Furthermore, the three components and various interactions between these provide starting points for interventions in the system that address possibilities for arbitrary conduct.

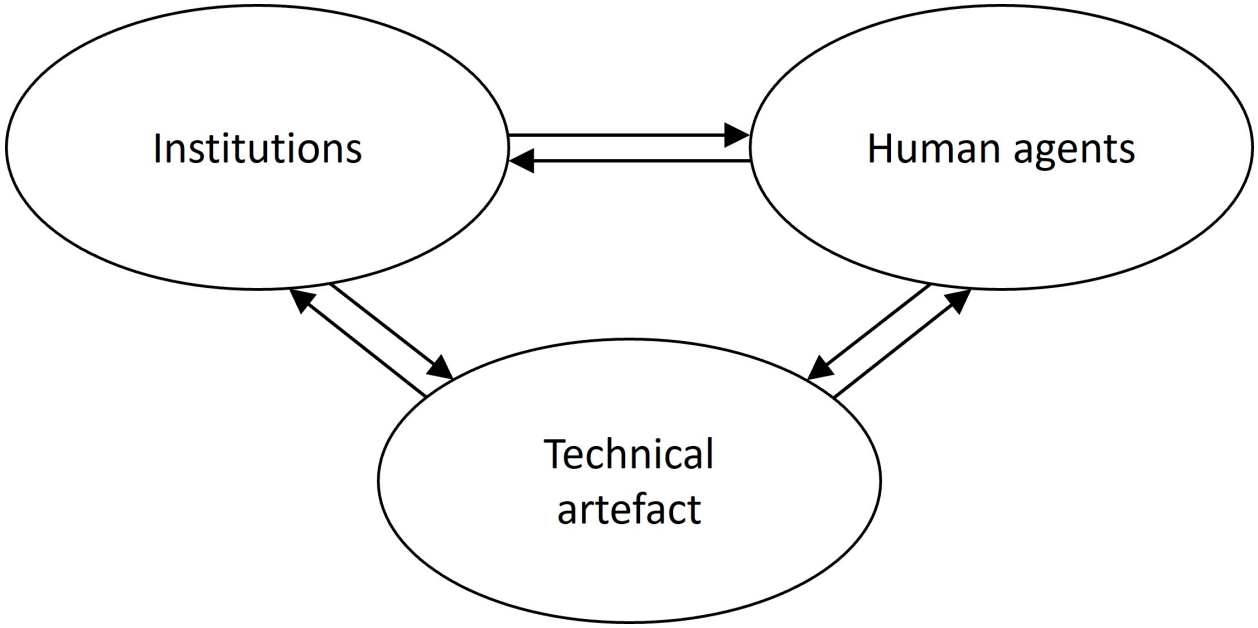

Figure 9.3 The socio-technical perspective comprises three components that structure each other [79]

Regarding arbitrary conduct from a socio-technical perspective provides a systematic appraisal of challenges and opportunities for the Rule of Law in cases of increased automation and augmentation of decision-making. First, there is the general threat to Rule of Law discussed by legal scholars.[80] Fig. 9.3 indicates that this threat actually refers to the technical component partly or fully replacing the function of institutions in public administration practices. In this case, the Rule of Law is sidelined by the interplay between the agential and technical component.[81] Second, the figure shows what the Rule of Law as an institutional endeavour can influence: both human behaviour as well as the form and function of the technical component within it. Finally, the figure points out the limitations of the Rule of Law's influence. It is always dependent on the technical artefact and the way in which human agents use both the technical as well as the institutional artefacts. Consequently, actors giving form to the Rule of Law are dependent on actors from other disciplines, e.g., developers of the AI application. Still, the Rule

[79] Figure by authors, based on Orlikowski 1992.
[80] E.g., Bayamlıoğlu and Leenes 2018; Hildebrandt 2018; Greenstein 2022.
[81] E.g., Endicott and Yeung 2022.

of Law can provide inspiration for the specification that shapes the technical artefact, and, based on the above, one may argue that the development practices of consequential AI applications as well as involved actors should fall under the Rule of Law.

9.4 The Rule of Law and system safety

This section discusses the starting points for a lexicon, an analysis, and a design practice based on system safety that can operationalise and ensure a socio-technical approach for the Rule of Law in public AI systems. First, the merits of a socio-technical perspective as presented in Sect. 9.3.3 are identified. Thereafter, we examine the elements that are missing in Rule of Law thinking. Finally, system safety is presented as a suitable perspective to enrich the socio-legal perspective on the Rule of Law.

9.4.1 Contribution of a socio-technical perspective to the Rule of Law

Bringing a socio-technical perspective to the Rule of Law has at least three contributions, all addressing overlooked systemic issues or flaws that contribute to arbitrary conduct. First, without a socio-technical understanding of failures in AI systems, the Rule of Law will retain its focus on illegitimate decision in individual cases, arbitrary conduct by individuals or instituting the Rule of Law on the state level. For example, individual citizens can go to court to contest a decision based on a public AI system. In case the decision is deemed erroneous, the individual decision is corrected or its consequences are compensated. However, this does not provide insight on whether this erroneous decision follows from a systemic flaw in the public AI system. König & Wenzelburger show how a design choice that determines the distribution of outcomes of AI systems can lead to arbitrary outcomes: whether a specific citizen is selected by the AI model does not depend on their own characteristics or applicability but on how these factors differ from other citizens in the population or the dataset.[82]

Second, the consequences of systemic interactions can be overlooked without a socio-technical perspective. Public AI systems may interact with other processes and systems, e.g. in how the outputs form inputs to other decisions, either within or outside the organisation it is situated in. These interactions may cause emerging effects in other, but related systems - also called *ripple effects* – that were not anticipated at the level of single processes or organizations,

[82] König and Wenzelburger 2021.

and which may significantly harm citizens.[83] Both sustaining the individual focus and neglecting systemic interactions may obscure responsibilities.

Finally, without acknowledging the socio-technical nature of public AI systems, the separation between Rule of Law practices and AI application design practices will persist and communication will be hindered. Following from the classic Rule of Law perspectives, it is sufficient to have law-makers craft legal institutions. The socio-legal perspective already shows that the behaviour of policymakers and users should be considered. On the other hand, AI applications are generally designed by technical developers who are guided by policies. The socio-technical perspective, therefore, puts emphasis on designing institutions and technology symbiotically.[84] This requires collaboration between actors from different disciplines,[85] and striving for alignment between institutional and technical components.[86]

The DUO case in the Netherlands shows the consequences of missing a socio-technical perspective. Until 2015, Dutch students were paid monthly study grants to cover a.o. costs of living by DUO – an executive agency of the ministry of education. Students living with their parents received a lower amount than students that lived on their own. This was determined on basis of address registration of students. A public AI system was implemented to detect fraud with the grants. The algorithm made a first selection of potential fraudsters that was refined by five public servants of DUO. Based on this selection, external bureaus checked the suspicion by paying house visits. The decision that followed could be contested at DUO first and, if denied, students could challenge the decision in court.[87]

In the summer of 2023, the discriminatory working of the public AI systems surfaced. Lawyers were noticing that a high number of students accused of fraud had different ethnic backgrounds. Moreover, in court, most of these accusations were annulled. Although there were several flaws in the institutional design such as the illegitimacy of the house visits by external parties, the case mostly shows that the algorithm was biased and that this bias was not resolved by the human-in-the-loop – i.e., the public servants of DUO. DUO did not evaluate its system on

[83] Peeters and Widlak 2018; Pel 2022.
[84] Koppenjan and Groenewegen 2005.
[85] De Bruijn and Herder 2009.
[86] Kunneke et al. 2021.
[87] Belleman B, Heilbron B, Kootstra A (2023) “Ik wil dat iemand zegt dat ik geen fraudeur ben”. De discriminerende fraudecontroles van DUO. https://www.platform-investico.nl/onderzoeken/de-discriminerende-fraudecontroles-van-duo Accessed 16 February 2024.

bias. [88] It took quite some time and effort from lawyers and journalists to detect this system flaw. The DUO case shows the lack of system level correction by only detecting and correcting individual cases.

9.4.2 Expanding perspectives on the Rule of Law

The problems discussed above show that the lexicon in the Rule of Law discipline falls short in examining socio-technical systems. This starts with the lack of a clear definition of what arbitrary conduct is.[89] The role of technical artifacts in arbitrary conduct is overlooked even more in currently available definitions. The lexicon emphasises institutions, mostly legal institutions, and the role of human conduct and social context. In that respect, the socio-legal perspective provides an important contribution by broadening the understanding of institutions. But for a full understanding of arbitrary conduct in public AI systems the role of the technical component also needs to be included in the Rule of Law lexicon.

Following from the deficiencies in the Rule of Law lexicon, the discipline lacks a methodology to detect where and analyse how public AI systems fall short in protecting citizens from harm. Again, in Rule of Law thought, analysis of arbitrary conduct focuses on what human behaviour counts as arbitrary use of power, how institutions are falling short to prevent or reduce such behaviour, and what institutions are needed to achieve that reduction. There is a need for an analytic approach that provides a comprehensive examination of opportunities for arbitrary conduct in public AI systems.

Similarly, Rule of Law thought provides too little grounding for the governance of public AI systems. Only intervening on the institutional level of public AI systems will fall short in addressing the role of the technical component in arbitrary conduct. As the technical component can play a dominant role in arbitrary conduct, there is a need for aligning technical and institutional design. Our approach will provide a first-of-its kind attempt towards bringing the design and use of public AI systems under the Rule of Law. We do this by complementing a socio-legal perspective on the Rule of Law by system safety.

[88] Ibid.
[89] Krygier 2016.

9.4.3 System safety

We argue that system safety can play a role addressing the three challenges of current Rule of Law approaches related to AI systems. System safety is a discipline that captures a long history of understanding harms and unsafe outcomes and, therefore, the protection of individuals.[90] The discipline initially arose in the aerospace and aviation domain and has since shaped safety standards and practices in various other sectors. Recently, lessons from system safety were resurfaced for common day AI applications and applied to a systematic study of algorithmic harms in the context of benefit allocation in The Netherlands.[91]

System safety goes beyond a technical perspective on systems. The discipline provides ways to address undesirable outcomes and harms of processes subject to forms of software-based automation. It treats such outcomes and harms as fundamentally 'emergent', which means that you can only understand these through considering social, technical and institutional components of the overall system, including the process of decision-making with its actual technological components, as well as the broader organization and institutional context. Similarly, the tools introduced by system safety address both the socio-technical specification of systems and the systems' working in practice. The discipline argues that technical as well as social and institutional interventions in the system are needed to ensure safety.

System safety does not impose safety as the main or only value that should be pursued. Instead, it provides a vocabulary and toolbox that serves as a system perspective for analysis and design to be used in different disciplines and domains. The lexicon and toolbox equip actors with a framework to map a particular system and to identify points of interventions in order to safeguard a selected value. This value does not necessarily have to be safety. In this chapter, we use the system safety perspective to strengthen values related to the Rule of Law – i.e., protecting citizens against arbitrary conduct. In other words, we are contextualising system safety by applying it to public AI systems and the Rule of Law.

Therefore, we bring the socio-legal perspective in conversation with the socio-technical perspective of system safety. The socio-legal perspective fits best because of its broad understanding of institutions. The combination of these perspectives can enhance the approach towards arbitrary conduct in three ways: a more comprehensive lexicon, an analytical lens that

[90] Leveson 2012.
[91] Dobbe 2022; Pel 2022.

also considers the technical component in arbitrary conduct, and a design approach for operationalizing Rule of Law principles that expands the current practice of institutional design from human conduct to including the form and function of AI applications in conduct. We will discuss these three ways in the next three sections.

9.5 Arbitrary conduct from a system perspective

System safety provides a lexicon to examine flaws in socio-technical systems that can lead to harms and other forms of damages or undesirable 'loss'. The lexicon hinges on the distinctive feature of system safety compared to traditional perspectives on safety as it focuses on hazardous system states instead of individual accidents. The lexicon pertains to system safety's socio-technical perspective that considers safety as an emergent property. This lexicon enables actors to investigate and communicate the internal working of systems that can lead to hazards and eventually to individual harms.

The lexicon is the starting point for the analysis and design efforts in system safety discussed in Sect. 9.6 and 9.7. In this section, we show that the lexicon can also provide useful insights for the Rule of Law into systemic issues in AI systems – i.e., we make the connection between system safety and the Rule of Law through the former's lexicon.

9.5.1 System safety lexicon

The main concept in system safety – *system hazard* – is defined as a 'system state or set of conditions that, together with a particular set of worst-case environmental conditions, will lead to an accident (loss).'[92] The system state here refers to a process in which a 'controller' makes decisions to arrive at outcomes that accomplish a certain objective and adhere to key (safety) constraints. This controller may be a human decision-maker, a group of human decision-makers, an automated decision-making entity, but can also be human decisions informed by or partly mediated by algorithms or information systems. The safety constraints determine what system behaviour is considered to be a hazard and, therefore, should be avoided through imposing these constraints.[93]

[92] Leveson 2012, p. 184
[93] Leveson 2012

Safety is considered to be a control problem in system safety. This means that safety constraints are enforced through control mechanisms. These control mechanisms play primarily at the *operational level*, meaning they form a key part of the actions and decisions that are available in the operational process in which the safety constraints need to be enacted. How a controller acts in the operational process is captured in a *process model*, which may be 'both embedded in the control logic of an automated controller or in the mental model maintained by a human controller.'[94]

Beyond the operational process are a variety of *secondary processes*, in which other control mechanisms may be needed to provide the right conditions for the enactment of safety procedures in operation. These processes are largely divided into system operations and system development. System operations includes the primary operational process, but also secondary processes such as operations management, company management, and other processes related to the operations of the main process including standard setting, oversight, insurance, courts, and legislation. System development includes maintenance, manufacturing, design, implementation, project management, innovation management, research, company management, and other processes related to the development of technologies used in the main process. It is thus understood and acknowledged that effective governance for safety requires control mechanisms and contributions from across these many processes and associated actors. The control mechanisms associated vary from audits to standards to work procedures, certification or legal penalties and much more. Each mechanism typically includes some *reference* that it is imposing on a process, which also requires a *measurement* to verify whether the reference is being met. All of the control mechanisms, both in the operational and secondary processes, that contribute to the enactment of safety, as well as the ways in which control mechanisms impose constraints on processes, are captured in a *safety control structure*.[95] Crucially, to ensure that the structural elements in system safety actually function, management and leadership are crucial, including the informal role of safety culture. This culture refers to the need to ensure that people critical in safeguarding a system feel safe themselves to raise possible safety issues, without the fear of

[94] Leveson 2012, p. 87
[95] Leveson 2012.

being retaliated against or others suffering from it, and that active follow-up of raised issues is organized.[96]

Finally, it is fundamental to define safety as an *emergent system property.* This means that *unsafe situations* can only be understood through the interactions between system components, including technical, human and institutional aspects. The canonical example to explain this is the fact that a valve can never be safe on itself. It is only safe in relation to the plant in which it is used and given that a plant operator knows when and how to open or close it to ensure that the plant runs safely. The emergent nature of systems also means that systems need to adapt to changing conditions and new insights throughout their life cycle. In a system's life cycle things happen such as, changing environmental conditions, and users that assign new functions to institutions and technical artefact. These changes can have an effect on safety in a system that those responsible have to address. A process of feedback to system designers on system hazards and adaptation of control actions is needed to address such changes. In general, achieving safety in socio-technical systems is a learning process.[97]

9.5.2 Connection to Rule of Law

The lexicon of system safety can be used to discuss arbitrary conduct in which an AI application is involved. The lexicon of system safety has commonalities with that of the Rule of Law. Both start from discussing the protection of individuals from harm. In this, they go back to the causes of such harms in either socio-legal or socio-technical systems. Still, system safety will help to reframe the concept of arbitrary conduct as it is explicitly expanded with the notions of the technical component.

We argue that the arbitrary use of power mediated by AI applications can also be considered an undesired emergent outcome. Part of the opportunities for arbitrary conduct only arise or materialise because of the interaction between system components. Especially over time, this can bring unexpected, unintended, or undesired situations of arbitrary conduct. Hence, we treat possibilities for arbitrary conduct as a subset of hazards with possible citizen harm or undesirable power imbalances as the resulting loss. This means that arbitrary conduct can be conceptualised as a system state, a set of environmental conditions, and control actions. Moreover, as for all

[96] Dekker 2012.
[97] Leveson 2012.

socio-technical systems, public AI systems need feedback channels; especially for detecting arbitrary conduct that is ingrained in the system and correcting for it at the appropriate system level (be it technical, social, or institutional).

9.6 System safety to analyse emergence of arbitrary conduct

System safety provides tools to analyse system hazards, where they emerge in current systems and what can go wrong in future systems. Considering arbitrary conduct as a system hazard, the analytical techniques of system safety can support in detecting possibilities for arbitrary conduct by tracking and tracing inadequate control mechanisms for arbitrary conduct. After detecting system hazards, system safety focuses on correcting the inadequate control mechanisms. This will be discussed in Sect. 9.7.

9.6.1 System safety analysis

Leveson suggests an analysis that focuses on two things: the controlled operational process (including process models) and the hierarchical safety control structure,[98] discussed in Sect. 9.5.1. Across these two, system safety analysis centres around the ex ante and ex post identification of *potential for inadequate control*. Inadequate control may happen at the level of actions in the operational process or at the level of institutional mechanisms in the safety control structure. Control mechanisms may either be (1) incorrect/unsafe, (2) not provided when necessary, (3) provided at the wrong time, or (4) applied too long or stopped too soon.[99] Each of such inadequacies may contribute to the system entering some unsafe state, violating a safety constraint. Analysis may start with the identification of inadequate control. It can also focus on a particular form of loss or harm, and try to understand what safety constraints are missing or how existing constraints were violated and what inadequacies in control actions and institutional mechanisms may contribute to it. Here, system safety both addresses direct causal scenarios but also *constitutive factors* (both organizational and systemic) that allow hazards to emerge in the operational process. Based on the analysis, concrete recommendations can be made to inform subsequent policy or system design efforts.[100]

[98] Leveson 2012.
[99] Ibid.
[100] Ibid.

9.6.2 Connection to Rule of Law

Mapping a public AI system through a system safety lens enables public organisations to detect potential for arbitrary conduct. The focus of this analysis is already defined in Rule of Law literature. Taekema suggests a 'focus on [the] locus of power and the opportunities for arbitrary conduct with adverse consequences for the interests of others.'[101]

The approach discussed in Sect. 9.6.1 support in tracing the three focus points suggested by Taekema: the locus of power, adverse consequences, and opportunities for arbitrary conduct. First, the system is mapped to determine the locus of power. The socio-technical specifications and practices in public AI systems are mapped by using the safety control structure, process models, and safety constraints. The safety control structure will also bring into view the design and maintenance processes needed to build and govern the technical artefacts. Analysts should consider at least four dimension in the case of public AI systems: (1) does the system comprise one or more processes?; (2) if there are more processes, what processes include an AI application?; (3) where in the system specification, system instantiation or human behaviour are possibilities for arbitrariness situated?; and (4) if these possibilities are emerging from human behaviour, what is the influence of design choices or the output of the system? Together, these dimensions will help identify the relevant loci of power in arbitrary conduct. Second, adverse consequences and opportunities for arbitrary conduct are identified by examining inadequate control actions and their causes.

9.7 Addressing arbitrary conduct through system safety

Possibilities for arbitrary conduct in public AI systems need to be prevented, mitigated, or corrected. System safety provides safety-guided design approaches to iteratively design required interventions into AI applications and the processes where these are used in. After detecting possible arbitrariness, the socio-technical specification and AI practices can be (re)designed following insights and methods from system safety. In the spirit of convention, this could inform 'Rule-of-Law'-Guided Design approaches currently emerging. In this section, we discuss the basic steps involved in such efforts.

[101] Taekema 2021, p. 90.

9.7.1 System safety as design approach

Once key hazards and their constitutive factors are in view through an ex ante or ex post analysis, these can be addressed through safety-guided design, which considers measures that are either technical, social, or institutional, or determine how these should interact. Once a new design is determined, additional hazard analysis can be done to determine whether the hazard and its constitutive factors have been addressed and no new hazards have emerged, hence the iterative nature.

In addressing hazards, the involved designers should follow a three-staged strategy.[102] First, they should see if the hazard can be fully eliminated, which means that the associated situation cannot happen anymore. Second, if elimination is not feasible or leads to disproportionate trade-offs, a next strategy is to reduce the likelihood of the hazard occurring, through installing new control mechanisms. Control mechanisms can either be *passive or active*. Passive controls provide safety constraints through their presence (e.g. a guard rail between a road and an adjacent abyss) and active controls need to be enacted through a safety mechanism that relies on taking control actions (e.g. a lane-keeping function to steer a car away veering into the abyss). The active control may be more functional, but also brings new vulnerabilities and hazards (e.g. the lane-keeping function may not work in certain weather conditions), which is why passive controls are often preferred (a guard rail works regardless of the weather). Third, if controls are not feasible or foolproof, designers should try to minimize the damage in the event a hazard leads to an accident. Efforts could include ways to dampen the impact (e.g. crumple zones and airbags) or to provide the operator or surrounding actors ways to minimize damage (e.g. providing an emergency hammer to leave a submerged car).

Once the various measures are designed and tested, one can draw up a *gap analysis*, which provides insight into what is needed to go from the current to the desired situation. Many different aspects may come into view, including the impact on production, use and maintenance of the system, the need for regulatory backing, the economic investments needed, and the desired level of knowledge and capability for those using, otherwise interacting with, or governing the system. Clearly, various trade-offs and political choices may emerge in this process, which allow decision-makers to properly gauge responsibilities and efforts to realize or transition towards the desired system configuration.

[102] Leveson 2012.

9.7.2 Connection to Rule of Law

The above insights help to identify and design measures that should support the legal institutions of the Rule of Law in contexts where AI is used in public administration processes. The socio-legal perspective prescribes two steps in arriving at measures:[103] (1) determine the immanent goal of the Rule of Law – i.e., reducing or addressing arbitrary use of power, and (2) design and implement measures that advance that immanent goal or purpose – adapted to the applicable context.[104] The first step we already covered in Sect. 9.6. The design approach of system safety helps to determine what interventions need to be designed.

In this design step, designers can also get inspiration from the classic perspective on legal institutions to address system hazards. Concerning designing hazards *out* of the system, the formal perspective on the Rule of Law can help. It lists formal requirements for rules that also apply to AI applications.[105] In this way, the formal requirements are grounded quality measurements for AI applications. The procedural perspective on the Rule of Law can provide ideas for institutional mechanisms that comprise the safety control structure. If an AI application is hazardous, its institutional environment needs meaningful recourse procedures and feedback mechanisms that citizens can instigate. A safety-guided approach to designing these necessarily includes ways to validate the efficacy of the control mechanism, both in terms of the *reference* it places as well as on the information or *measurement* that is taken to verify whether the reference is met. Crucially, these procedures and feedback mechanisms should provide the possibility to assign a public AI system back to its design mode. For example, front workers should have the room to question the output of an AI application, and have it be changed if the output or the logic through which it is produced has significant flaws or does not meet the needs of the context. Not being able to adjust the technical workings of the AI system often means that the issue persists, and worse, that servants lose discretion and citizens their ability to contest, being more prone to be dependent on post-hoc legal protection, if available at all.

[103] Krygier 2009.
[104] Selznick 1999.
[105] Brownsword 2016.

9.8 Conclusions

As they are intervening in the relationship between citizen and state, public AI systems and their design processes fall under the Rule of Law regime. AI systems can mediate arbitrary conduct and their socio-technical nature requires adapted approaches towards translating the Rule of Law to AI practices. We argue that this can be done by combining the socio-legal perspective on the Rule of Law with a system safety perspective on AI systems. This synthesis implies that the analytical lens of system safety can support the detection of possibilities for arbitrariness in public AI systems. Moreover, it can support the translation of Rule of Law principles to the socio-technical specification of public AI systems and their associated design and use practices. Applying this combined perspective will not be straightforward. The influence of AI practices on public administration is still highly uncertain, the capacities of public organisation are premature, and these practices cross conventional boundaries and are highly contextual. Nevertheless, the altering of power dynamics in public administration by AI practices urges for protection of citizens against arbitrary use of power by the state mediated by public AI systems.

References

Alkhatib A (2021) To Live in Their Utopia: Why Algorithmic Systems Create Absurd Outcomes. Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems. CHI '21: CHI Conference on Human Factors in Computing Systems DOI:10.1145/3411764.3445740

Alkhatib A, Bernstein M (2019) Street-Level Algorithms: A Theory at the Gaps Between Policy and Decisions. Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. CHI '19: CHI Conference on Human Factors in Computing Systems DOI:10.1145/3290605.3300760

Barocas S, Selbst AD (2016) Big Data's Disparate Impact. DOI:10.15779/Z38BG31

Bayamlıoğlu E, Leenes R (2018) 'The "rule of law" implications of data-driven decision-making: a techno-regulatory perspective. Law, Innovation and Technology DOI:10.1080/17579961.2018.1527475

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 Administration Review DOI:10.1111/0033-3352.00168

Braithwaite V (2020) Beyond the bubble that is Robodebt: How governments that lose integrity threaten democracy. Australian Journal of Social Issues DOI:10.1002/ajs4.122

Brownsword R (2016) Technological management and the Rule of Law. Law, Innovation and Technology DOI:10.1080/17579961.2016.1161891

Burrell J (2016) How the machine "thinks": Understanding opacity in machine learning algorithms. Big Data & Society DOI:10.1177/2053951715622512

Chiao V (2023) Algorithmic Decision-Making, Statistical Evidence and the Rule of Law. Episteme DOI:10.1017/epi.2023.27

Cobbe J, Veale M Singh J (2023) Understanding accountability in algorithmic supply chains. 2023 ACM Conference on Fairness, Accountability, and Transparency. FAccT '23: the 2023 ACM Conference on Fairness, Accountability, and Transparency DOI:10.1145/3593013.3594073.

Cuéllar M, Huq AZ (2022) Artificially Intelligent Regulation. Daedalus DOI:10.1162/DAED_a_01920

De Bruijn H, Herder PM (2009) System and Actor Perspectives on Sociotechnical Systems. IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans DOI:10.1109/TSMCA.2009.2025452

De Bruijn H, Warnier M, Janssen M (2022) The perils and pitfalls of explainable AI: Strategies for explaining algorithmic decision-making. Government Information Quarterly DOI:10.1016/j.giq.2021.101666

Dekker S (2012) Just culture: balancing safety and accountability, 2nd edn. Ashgate, Aldershot

Dobbe R, Dean S, Gilbert T, Kohli N (2018) A Broader View on Bias in Automated Decision-Making: Reflecting on Epistemology and Dynamics. http://arxiv.org/abs/1807.00553. Accessed 16 February 2024

Dobbe R, Krendl Gilbert T, Mintz Y (2021) Hard choices in artificial intelligence. Artificial Intelligence DOI:10.1016/j.artint.2021.103555

Dobbe RIJ (2022) System Safety and Artificial Intelligence. The Oxford Handbook of AI Governance DOI:10.1093/oxfordhb/9780197579329.013.67.

Endicott T, Yeung K (2022) The death of law? Computationally personalized norms and the rule of law. University of Toronto Law Journal DOI:10.3138/utlj-2021-0011.

Fountain JE (2001) Building the virtual state: information technology and institutional change. Brookings Institution Press, Washington D.C.

Giddens A (1976) New Rules of Sociological Method. Basic Books, New York

Giddens A (1979) Central Problems in Social Theory: Action, Structure and Contradiction in Social Analysis. University of California Press, Berkeley

Giddens A (1984) The Constitution of Society: Outline of the Theory of Structure. University of California Press, Berkeley

Green B (2022) The flaws of policies requiring human oversight of government algorithms. Computer Law & Security Review DOI:10.1016/j.clsr.2022.105681

Green B, Chen Y (2019) Disparate Interactions: An Algorithm-in-the-Loop Analysis of Fairness in Risk Assessments. Proceedings of the Conference on Fairness, Accountability, and Transparency. FAT* '19: Conference on Fairness, Accountability, and Transparency DOI:10.1145/3287560.3287563.

Greenstein S (2022) Preserving the rule of law in the era of artificial intelligence (AI). Artificial Intelligence and Law DOI:10.1007/s10506-021-09294-4

Grimmelikhuijsen S, Meijer A (2022) Legitimacy of Algorithmic Decision-Making: Six Threats and the Need for a Calibrated Institutional Response. Perspectives on Public Management and Governance DOI:10.1093/ppmgov/gvac008.

Hildebrandt M (2011) Legal Protection by Design: Objections and Refutations. Legisprudence DOI:10.5235/175214611797885693.

Hildebrandt M (2016) Law as Information in the Era of Data-Driven Agency. The Modern Law Review DOI:10.1111/1468-2230.12165.

Hildebrandt M (2018) Algorithmic regulation and the rule of law. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences DOI:10.1098/rsta.2017.0355.

Hildebrandt M (2019) Privacy as Protection of the Incomputable Self: From Agnostic to Agonistic Machine Learning. Theoretical Inquiries in Law DOI:10.1515/til-2019-0004.

Hodgson GH (2006) What are institutions? Journal of Economic Issues DOI:10.1080/00213624.2006.11506879

Janssen M, Kuk G (2016) The challenges and limits of big data algorithms in technocratic governance. Government Information Quarterly DOI:10.1016/j.giq.2016.08.011

König PD, Wenzelburger G (2021) The legitimacy gap of algorithmic decision-making in the public sector: Why it arises and how to address it. Technology in Society DOI:10.1016/j.techsoc.2021.101688

Koppenjan J, Groenewegen J (2005) Institutional design for complex technological systems. International Journal of Technology, Policy and Management DOI:10.1504/IJTPM.2005.008406

Krygier M (2009) The Rule of Law: Legality, Teleology, Sociology. Relocating the Rule of Law DOI:10.5040/9781472564634

Krygier M (2014) Rule of Law (and Rechtsstaat). The Legal Doctrines of the Rule of Law and the Legal State (Rechtsstaat) DOI:10.1007/978-3-319-05585-5_4

Krygier M (2016) The Rule of Law: Pasts, Presents, and Two Possible Futures. Annual Review of Law and Social Science DOI:10.1146/annurev-lawsocsci-102612-134103

Kunneke R, Ménard C, Groenewegen J (2021) Network Infrastructures: Technology meets Institutions. Cambridge University Press

Leveson NG (2012) Engineering a Safer World: Systems Thinking Applied to Safety. The MIT Press

Mak E, Taekema S (2016) The European Union's Rule of Law Agenda: Identifying Its Core and Contextualizing Its Application. Hague Journal on the Rule of Law DOI:10.1007/s40803-016-0022-1

Mulligan DK, Bamberger KA (2019) Procurement as Policy: Administrative Process for Machine Learning'. DOI:10.15779/Z38RN30793

Nissenbaum H (2019) Contextual Integrity Up and Down the Data Food Chain. Theoretical Inquiries in Law DOI:10.1515/til-2019-0008

Nonet P, Selznick P (2001) Law & society in transition: toward responsive law. Transaction Publishers, New Brunswick, N.J.

Orlikowski WJ (1992) The Duality of Technology: Rethinking the Concept of Technology in Organizations. Organization Science DOI:10.1287/orsc.3.3.398

Passchier R (2020) Digitalisering en de (dis)balans binnen de trias politica [Digitalisation and the (dis)balance within the trias politica]. Ars Aequi 69(10): 916–927

Peeters R, Widlak A (2018) The digital cage: Administrative exclusion through information architecture – The case of the Dutch civil registry's master data management system. Government Information Quarterly DOI:10.1016/j.giq.2018.02.003

Peeters R, Widlak AC (2023) Administrative exclusion in the infrastructure-level bureaucracy: The case of the Dutch daycare benefit scandal. Public Administration Review DOI:10.1111/puar.13615

Pel L (2022) Ripple Effects of Law Execution Automation in Governmental Systems: The Wajong Case. Delft University http://resolver.tudelft.nl/uuid:64271d5f-b3b3-4951-ba13-4c6c9435d9ec. Accessed 16 February 2024

Raz J (1979) The Rule of Law and its Virtue. The authority of law: Essays on law and morality DOI:10.1093/acprof:oso/9780198253457.003.0011

Santoni De Sio F, Mecacci G (2021) Four Responsibility Gaps with Artificial Intelligence: Why they Matter and How to Address them. Philosophy & Technology DOI:10.1007/s13347-021-00450-x

Sculley D, Holt G, Golovin D, Davydov E, Phillips T, Ebner D, Chaudhary V, Young M, Crespo J, Dennison D (2015) Hidden Technical Debt in Machine Learning Systems. NIPS DOI:10.5555/2969442.2969519

Seaver N (2017) Algorithms as culture: Some tactics for the ethnography of algorithmic systems. Big Data & Society DOI:10.1177/2053951717738104

Selznick P (1999) Legal Cultures and the Rule of Law. The Rule of Law after Communism: Problems and Prospects in East-Central Europe DOI:10.4324/9781315085319

Selznick P (2003) "Law in Context" Revisited. Journal of Law and Society DOI:10.1111/1467-6478.00252

Taekema S (2021) Commitment to the Rule of Law: From a Political to an Organizational Ideal. Ethical Leadership in International Organizations DOI:10.1017/9781108641715.003

Veale M, Brass I (2019) Administration by Algorithm?: Public Management Meets Public Sector Machine Learning. Algorithmic Regulation DOI:10.1093/oso/9780198838494.003.0006

Waldron J (2011) The rule of law and the importance of procedure. Getting to the Rule of Law DOI: 10.18574/nyu/9780814728437.003.0001

Whittaker, M. *et al.* (2018) AI Now Report 2018. AI Now Institute.

Whittaker M (2021) The steep cost of capture. Interactions DOI:10.1145/3488666

Widder DG, Nafus D (2023) Dislocated accountabilities in the "AI supply chain": Modularity and developers' notions of responsibility. Big Data & Society DOI:10.1177/20539517231177620

Widlak A, Peeters R (2020) Administrative errors and the burden of correction and consequence: how information technology exacerbates the consequences of bureaucratic mistakes for citizens. International Journal of Electronic Governance DOI:10.1504/IJEG.2020.106998

Yeung K (2014) Design for the Value of Regulation. Handbook of Ethics, Values, and Technological Design DOI:10.1007/978-94-007-6994-6_32-1

Yeung K (2018) Algorithmic regulation: A critical interrogation. Regulation & Governance DOI:10.1111/rego.12158

Young MM, Himmelreich J, Bullock JB, Kim, K (2021) Artificial Intelligence and Administrative Evil. Perspectives on Public Management and Governance DOI:10.1093/ppmgov/gvab006

Zalnieriute M, Moses LB, Williams G (2019) The Rule of Law and Automation of Government Decision-Making. The Modern Law Review DOI:10.1111/1468-2230.12412

Zouridis S, Van Eck M, Bovens M (2020) Automated Discretion. Discretion and the Quest for Controlled Freedom DOI:10.1007/978-3-030-19566-3_20.