Integral system safety for machine learning in the public sector

An empirical account

Delfos, J.; Zuiderwijk, A. M.G.; van Cranenburgh, S.; Chorus, C. G.; Dobbe, R. I.J.

DOI
10.1016/j.giq.2024.101963

Publication date
2024

Document Version
Final published version

Published in
Government Information Quarterly

Citation (APA)
Delfos, J., Zuiderwijk, A. M. G., van Cranenburgh, S., Chorus, C. G., & Dobbe, R. I. J. (2024). Integral system safety for machine learning in the public sector: An empirical account. *Government Information Quarterly*, *41*(3), Article 101963. https://doi.org/10.1016/j.giq.2024.101963

Government Information Quarterly 41 (2024) 101963

Contents lists available at ScienceDirect

Government Information Quarterly

journal homepage: www.elsevier.com/locate/govinf

Integral system safety for machine learning in the public sector: An empirical account

J. Delfos (Jeroen)*, A.M.G. Zuiderwijk (Anneke), S. van Cranenburgh (Sander), C.G. Chorus (Caspar), R.I.J. Dobbe (Roel)

Delft University of Technology, Faculty of Technology, Policy and Management, Jaffalaan 5, 2628, BX, Delft, the Netherlands

ARTICLE INFO

Keywords:
Machine learning
Artificial intelligence
Systems theory
System safety
Public sector
Public policy
Governance

ABSTRACT

This paper introduces systems theory and system safety concepts to ongoing academic debates about the safety of Machine Learning (ML) systems in the public sector. In particular, we analyze the risk factors of ML systems and their respective institutional context, which impact the ability to control such systems. We use interview data to abductively show what risk factors of such systems are present in public professionals' perceptions and what factors are expected based on systems theory but are missing. Based on the hypothesis that ML systems are best addressed with a systems theory lens, we argue that the missing factors deserve greater attention in ongoing efforts to address ML systems safety. These factors include the explication of safety goals and constraints, the inclusion of systemic factors in system design, the development of safety control structures, and the tendency of ML systems to migrate towards higher risk. Our observations support the hypothesis that ML systems can be best regarded through a systems theory lens. Therefore, we conclude that system safety concepts can be useful aids for policymakers who aim to improve ML system safety.

1. Introduction

Machine learning (ML) systems are increasingly used in the public sector (Engstrom et al., 2020; van Noordt & Misuraca, 2022). The expectation of such systems is that they make public services cheaper and more effective (Maciejewski, 2017). These benefits are expected from ML's ability to offer personalized services, make more accurate forecasts, and model complex systems (Margetts & Dorobantu, 2019). Examples of ML uses in the public sector can be found in the interaction with citizens (Aoki, 2020), detection of fraud (Pérez López et al., 2019), and profiling (Brennan et al., 2009). ML systems include a range of algorithms that learn from data and subsequently can predict new data (Zhou, 2021). It is the ability to learn from data that makes ML systems powerful tools for governments, who gather large amounts of data during their administrative duties.

Besides the opportunities, scholars have identified several negative implications of ML systems in the public sector. For example, the idea that ML systems might deliver biased results across different groups or individuals with particular characteristics is widely acknowledged (Fountain, 2022; Mehrabi et al., 2022). Biased outcomes of ML systems may result from biased data or design choices within the ML algorithm itself (Mehrabi et al., 2022), potentially leading to unfair and discriminatory decisions (Fountain, 2022; Kroll et al., 2015). Even if an ML system has little bias, it will still make errors, which can lead to incorrect decision-making and harmful outcomes (Dobbe, 2022). Additionally, working with ML systems can lead to unclarity in the attribution of responsibilities. The opacity of ML systems, found in, for example, artificial neural networks, can hinder civil servants from accessing coherent explanations for the outcomes of ML, thereby posing challenges to justifying and scrutinizing decisions (Janssen & Kuk, 2016; Wieringa, 2020). Furthermore, privacy breaches may be induced by the use of ML systems. The potential of ML systems may move organizations to use personal data unlawfully (Broeders et al., 2017). Although some of these negative implications are relevant for private organizations, ML system applications in the public sector present unique challenges (Desouza et al., 2020). These challenges include the public sector's need for transparency (Bryson & Winfield, 2017), the high diversity of stakeholders with conflicting agendas (Desouza et al., 2020), and the general requirement of ML systems in the public sector to promote the public good (Cath et al., 2018).

* Corresponding author.
E-mail addresses: j.delfos@tudelft.nl (J. Delfos), A.M.G.Zuiderwijk-vanEijk@tudelft.nl (A.M.G. Zuiderwijk), S.vanCranenburgh@tudelft.nl (S. van Cranenburgh), C.G.Chorus@tudelft.nl (C.G. Chorus), R.I.J.Dobbe@tudelft.nl (R.I.J. Dobbe).

https://doi.org/10.1016/j.giq.2024.101963
Received 6 October 2023; Received in revised form 26 July 2024; Accepted 15 August 2024
Available online 23 August 2024

Several policy and regulation initiatives have emerged on both national and international levels to counter these negative implications. First, there have been initiatives to guide public algorithmic system design with the help of guidelines (High-Level Expert Group on AI, 2019). Second, calls for transparency have led to policies requiring public organizations to be more transparent about using algorithmic systems (Artificial Intelligence Act, 2024; Overheid.nl, 2023). Third, supervisory agencies are tasked with supervising the public and commercial use of algorithmic systems (Artificial Intelligence Act, 2024; Dutch Data Protection Authority, 2023). Fourth, regulations are in the making that target both public and commercial use of Artificial Intelligence in Europe (Artificial Intelligence Act, 2024). These policies and regulations aim to improve the safety of ML systems. The first objective of the Artificial Intelligence Act is to "ensure that AI systems placed on the Union market and used are safe and respect existing law on fundamental rights and Union values" (Artificial Intelligence Act, 2024, p. 3).

There is an ongoing discussion about how policies should be designed to make ML systems safer. Within these discussions, a common critique of current practices for designing these policies can be observed: ML systems are part of complex socio-technical systems without a single and easy path towards safety. For example, first, addressing bias in the data is needed but not enough since the practices that lead to bias are embedded in broader societal problems (Fountain, 2022; Hoffmann, 2019). Second, the explainability of ML systems is helpful for detecting errors produced by these systems, but mistakes will still be made (Janssen et al., 2022). Third, placing an ML system in an existing system will change the behavior of actors in this system, which cannot be accounted for by the technological components of an ML system (Selbst et al., 2019).

However, we identify three major gaps in the literature that hinder the promotion of the safety of ML systems. First, ML systems are sparsely researched within the context in which they are implemented. When describing these systems' negative implications, there is a focus on the design phase and the ML system's technical artifacts. This focus covers only a part of the implications, as some only emerge when a system is implemented (Dobbe et al., 2018). The implementation of ML systems poses different challenges compared to the design phase. For example, it requires different expertise, which is found to be lacking, especially in the public sector (van Noordt & Tangi, 2023). Furthermore, the impact of ML systems is highly context-dependent and can only be assessed when considering institutions, processes (Gansky & McDonald, 2022), and social and political contexts (Cath & Jansen, 2022). Zuiderwijk et al. (2021) note that the context of the public sector has scarcely been the subject of research in conceptual and practice-driven studies. Although scholars have advocated for adopting non-technical and contextual factors in, for example, frameworks for data governance (Janssen et al., 2020) and strategies for explaining algorithmic decision-making (de Bruijn et al., 2022), this approach is lacking when studying the negative implications of ML systems in the public sector.

Second, we see a gap in applying rigorously defined concepts when discussing ML system implications. Risks, hazards, and challenges, amongst others, are terms commonly used to indicate negative implications but are not linked to concepts found in safety disciplines, some of which have intimate insight into software-based systems (Dobbe, 2022). Studies that map the challenges of ML systems have provided relevant overviews of aspects of ML systems (e.g., Sun & Medaglia, 2019; Wirtz et al., 2022), but it is left undescribed how these aspects relate to each other. For example, Wirtz et al. (2022) describe both 'discrimination of minorities' and 'defining human values' as 'ethical AI risks', without further explaining any causal or hierarchical relationship between these aspects. This lack of clear conceptualization hinders the progress towards a deeper understanding of the implications of ML systems in the public sector, as well as a proper diagnosis of causal relationships involved and possible interventions to prevent or mitigate such implications. Such understanding is crucial for developing effective policies, regulations, and organizational capabilities to promote the safety of ML systems.

Third, there is a gap in the empirical underpinning of conceptual studies into ML systems used in the public sector (Aoki, 2020; Zuiderwijk et al., 2021). Such Additional empirical data will provide a more comprehensive and evidence-based understanding of ML implementation's actual consequences and dynamics in real-world contexts.

In this paper, we aim to fill these gaps. The first gap, researching ML systems in their implemented context, is addressed by adopting a systems perspective. Such a systems perspective has been advocated for by several scholars (Dwivedi et al., 2021; Janssen & Kuk, 2016; Straub et al., 2023). We use systems theory as our base for this systems perspective. Systems theory has been the main underlying theory in the research field of system safety (Leveson, 2011). This research field has proven to be valuable in several engineering sectors, such as the space, aviation and automotive sector. Although conceptual studies have shown that valuable lessons can be drawn from system safety, its concepts and ideas have yet to find their way into the field of information systems. We use the work of Leveson (2011) on system safety, which introduces a set of rigorously defined concepts through which systems thinking can be used to address the safety of systems. In doing so, we address the lack of clear conceptualization of safety concepts. Lastly, we introduce new empirical data to address the third gap, conceptual work's lack of empirical underpinning. This data is gathered during interviews with twelve Dutch public professionals who represent the main stakeholders in the design, management, and supervision of ML systems for public decision-making. Filling these literature gaps results in a set of perceptions of the interviewees, which can be explained by systems theory. Furthermore, there are themes that, when adopting a systems theory lens, one would expect to observe in the perceptions of the interviewees but that were not found. We argue that these themes are in fact relevant and may be given more attention in the ongoing debate about, and the design of, policies for ML systems safety.

We argue that the analysis of perceptions of public professionals working in the field of ML systems through a systems theory lens is a novel contribution to the field of information systems. Furthermore, literature that couples the system safety discipline to the domain of information systems is still limited. This paper may serve as a starting point for further research, which uses concepts of system safety to address ML systems risks. This paper yields a first exploration of its merits. The chosen method for data collection, i.e., semi-structured interviews, is suited for the exploratory character of this study and the 'investigation of causation' (Gorman et al., 2005). Furthermore, this study makes a societal contribution to increase ML systems safety. Citizens are skeptical of the use of ML systems in the public sector (Haesevoets et al., 2024), while citizen trust in such systems is critical for its success (Wirtz et al., 2019). To change citizen attitudes towards ML systems, there is a need to move beyond the 'checkbox mentality' and search for ways to prevent incidents that violate citizen trust in the long term (Kleizen et al., 2023). We argue that a systems theory approach provides a base for ML system safety.

In the following sections, we introduce our theoretical and conceptual framework. We then explain our method, after which we present the perceptions of the interviewees and link these perceptions to system safety concepts in the results section. We proceed with a discussion in which we identify gaps between the perceptions of the interviewees and system safety and reflect on the applicability of systems theory for addressing ML systems in the public sector. We end the paper with the limitations and implications of our findings for further research and policy-making, and our final conclusions.

2. Theoretical and conceptual framework

Through abductive reasoning, we use systems theory and system safety concepts to explain observations about the risk factors of ML systems in the public sector. Systems theory and system safety form the conceptual framework used to analyze the interview data presented in

this paper. In this section, we first show why a systems theory lens is applicable for assessing ML systems safety. Subsequently, we introduce core concepts of system safety for which we use the work of Leveson (2011). These concepts will be used to abductively analyze the perceptions that we find in the interview data presented in this paper.

2.1. Systems theory and machine learning systems

Although Computer Science and Statistics are at the heart of ML systems, its impacts can only be understood in a broader context. The human designers of ML systems introduce values and biases in its technical design (Janssen & Kuk, 2016), and the data used to train the model reflect the human biases of those who registered this data (Fountain, 2022). Furthermore, after its design, ML systems are placed in a context involving human decision-makers interacting with the technical system (Dobbe et al., 2021). These decision-makers may or may not agree with ML systems output, which is, amongst others, dependent on their experience (Janssen et al., 2022). Factors such as the values of the human designers, bias in training data, and the experience of human decision-makers are thus key for understanding the workings of ML systems. Adopting a scope beyond the technical artifacts and adopting a systems perspective when looking at ML systems has been advocated by scholars (Janssen & Kuk, 2016; Straub et al., 2023).

System theory goes one step further than extending the scope beyond technical artifacts when analyzing a system. It states that systems with a certain degree of complexity must be analyzed as a whole and that the analysis of subsystems independently will not adequately describe the full system (Leveson, 2011). These systems show emergent properties that cannot be foreseen when looking at component level behavior of the system. ML systems are such systems in which not all behavior can be explained at the component level (Dobbe et al., 2021).

2.2. System safety

Systems theory has been used to develop the analysis of the safety of engineering systems. The field that originated from this line of thinking is called system safety. System safety has been employed to improve the safety of, for example, aircraft and spacecraft (Roland & Moriarty, 1990) and software systems (Leveson, 1995). One of the most influential viewpoints on system safety can be found in the work of Leveson (2011), which we will use for the remainder of this section to explain the concepts of system safety. The work of Leveson stand out in its potential for practitioners responsible for (policies for) the safety of systems. Tools for making systems safer and for analyzing accidents are adopted in several engineering sectors. Leveson and Weiss (2009) show the relevance of system safety for software, and more specifically, (Dobbe, 2022) shows that valuable lessons can be learned from system safety concepts for addressing the safety of Artificial Intelligence systems.

System safety deviates from more traditional safety engineering in seven ways. First, it defies the assumption that safety can be defined on a system component level. Rather, accidents result from the interaction between these components, even when these components are not failing individually. Second, it challenges the common notion that accidents result from 'event chains' that have a 'root cause,' with the observation that the entire socio-technical system contributes to safety. This includes organizational and social components such as organizational culture and safety policy. Third, it points out that probabilistic risk analysis omits risk factors that are difficult to quantify and that fixing these risk factors should be prioritized over measuring them. Fourth, it describes operator errors as a result of the environment the operator works in instead of seeing errors purely as the result of human failure. Fifth, it makes a clear distinction between safety and reliability. Systems can be safe but unreliable or unsafe but reliable. As such, what looks like functional and reliable software may be unsafe, and software complexity management is crucial in preventing additional safety risks. Sixth, it adopts the view that systems are dynamic and tend to become more unsafe over time as safety measures degrade. The constantly changing context in which a system operates means that today's safety measures may not be enough to ensure safety tomorrow. Seventh, it sees a danger in attributing blame in response to accidents, as this does not help the analysis of how the system behavior led to the accident. There is, hence, a natural tradeoff between accountability and safety, particularly at the level of organizational culture.

System safety uses a set of concepts to describe safety. We introduce these concepts in the remainder of this section. Fig. 1 shows a schematic representation of these concepts in cohesion. The ML system and its context determine the risk factors. These risk factors influence the ability to control the behavior of the system. Furthermore, the ML system and its context determine how safety management is shaped, which in turn determines the control strategies. Both the control strategies and the risk factors are the input for human and automated control. If the control is performed correctly and, thus, all control conditions are met, safety is achieved. If not, the system will enter a hazardous state and there is a probability, indicated by the dashed arrow, for an accident to occur.

Safety itself is regarded as the absence of accidents. It is an emergent property of a system, meaning it can only be evaluated in the context of the complete system and not the technical components alone. Instead, a system comprises technical and non-technical elements. To highlight the inclusion of non-technical aspects as a part of a system, the term sociotechnical systems is frequently used in system safety. The boundaries of the system can be chosen arbitrarily, but to assess the safety of the system, it should include those components over which the system designer has control. Accidents are "undesired or unplanned events that result in a loss." (Leveson, 2011, p. 181).

Accidents occur in certain system states in combination with worst-case environmental conditions, which are called 'hazards'. Identifying such hazards and their causal factors is called a 'hazard analysis'. We call these causal factors 'risk factors. Besides technical factors related to the soft- and hardware of systems, system safety stresses the importance of including non-technical factors such as norms, rules, and standards as well as organizational culture and safety policy. In this paper, we will refer to these factors as 'institutional factors', in line with Koppenjan and Groenewegen (2005), p. 244), who define institutions as "a set of rules that structure the course of actions that a set of actors may choose."

In system safety, safety is a 'control problem'. This means that the operational process, operated by humans, automation, or a mix/interaction thereof, aims to respect 'safety constraints' that limit the behavior of the system to safe situations. This includes preventing the system from being in 'hazardous situations,' which are situations from which, under certain events, an unsafe situation can emerge. Four conditions need to be met to have effective control for operators to steer away from hazards. First, it must be clear to the controller what the control goal and the safety constraints are. Second, the controller needs to be able to influence the system state to steer away from unsafe situations. Third, the controller must possess a model of the system containing information about the current state of the system, the relation between system variables, and how to influence the system. Fourth, the controller must be able to observe in what state the system is or might enter in some upcoming time window.

Control exists through observations of the system state and actions to influence the system state towards the desired system state. For the operation of the system, observations are done through sensors or data gathering (including human sensing), and the system state is influenced by actuators and/or control actions, either automated, manually operated, or a mix thereof. The system safety field describes the need for controls in a sociotechnical system, which may be enacted at varying levels, both in the operational process as well as in other neighboring processes, including but not limited to design, maintenance, management, or supervision. Every level of control has its 'reference channel' to impose a particular criterion on what is being controlled and a 'measuring channel' to understand to what extent the criteria are being met. The highest level of control in a sociotechnical system comes from

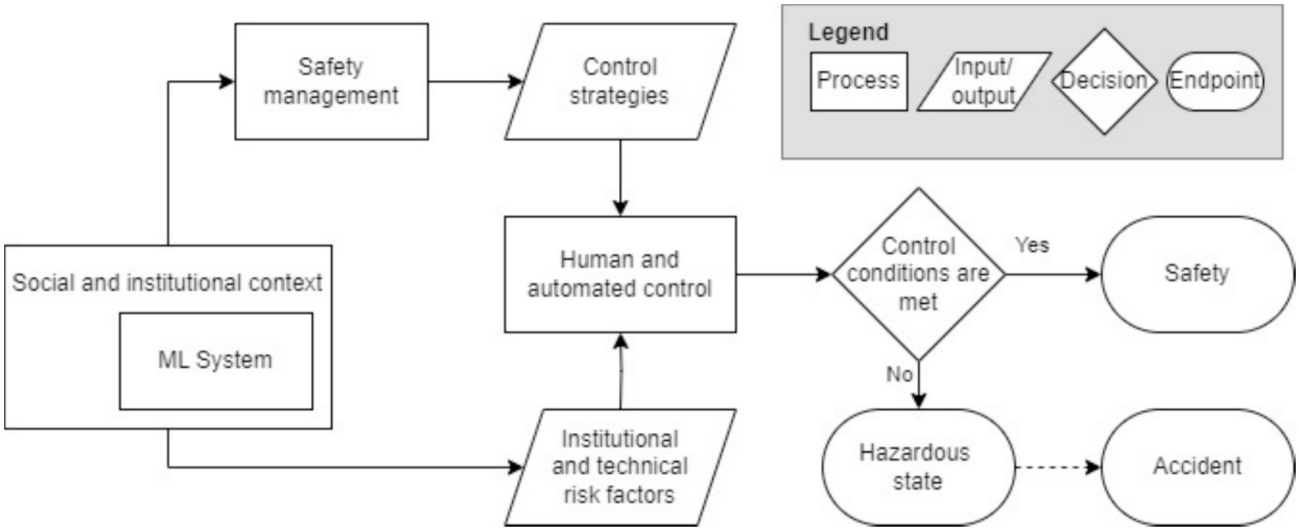

Fig. 1. Schematic representation of the relation between the System Safety concepts derived from Leveson (2011).

legislators who influence the system by means of legislation (reference channel) and observe the system state through, for example, oversight reports or accident investigations. Lower levels include regulatory agencies and management layers, which pass regulations, certification requirements, and standards. At the lowest level, we have the actual operational process in which the technology is applied and where various control measures are implemented to safeguard a system in real-time. In the remainder of this paper, we will refer to the options to control the sociotechnical system as 'control strategies.'

Creating and maintaining these control strategies is called 'safety management'. The objective of safety management is to maximize flexibility and improve system performance while adhering to safety constraints. System safety describes several ingredients for effective safety management, including commitment and leadership, a strong safety culture, and education and training.

3. Method

This section discusses the method for data collection and analysis.

3.1. Data collection

This study employs a qualitative data collection methodology, allowing us to explore the perceptions of risk factors of ML system safety. We interviewed Dutch professionals from executive organizations working with ML systems and from external supervisors who supervise the use of ML systems at public organizations. The Dutch context is particularly interesting because of the relatively high adoption of ML systems in the public sector (van Noordt & Misuraca, 2022).

Our sample includes public stakeholders who are closely involved with implemented ML systems and, therefore, have experience of how hazardous situations can arise. Based on systems theory, we view ML systems as hierarchical systems. The interviewees represent actors in the hierarchical layers of the control of ML systems that have a view of and say in the operation of the system (Leveson, 2011, p. 82). These actors include data scientists, operational managers, organization management, and external supervisors. For the interviewees in the category of external supervisors, both actors with deeper technical knowledge as well as managers were included. This group of actors from different hierarchical levels provides a wide spectrum of perspectives on the risk factors of ML systems in the public sector.

Twelve public professionals were interviewed, working at ten different organizations within the Dutch government. The last interviews yielded little new information regarding our research goal, after which no more interviews were conducted. We used a semi-structured approach that included a set of predefined questions, which are presented in Supplement 1. The open-ended questions left room for unexpected answers, which was suited for the exploratory character of this study and the 'investigation of causation' (Gorman et al., 2005). Although the interviewees can be regarded as experts in the field of ML usage in the public sector, they are not experts in system safety. As we aimed to explore possible risk factors, we asked the interviewees rather broadly to reflect on 'the risks' of ML in the public sector and possible risk 'mitigation strategies.' Explicitly asking to reflect on mitigation strategies proved to be helpful in getting a deeper understanding of the risk factors that the mitigation strategies would have to counter.

An overview of the interviewees can be found in Table 1. Note that interview I11 was an interview with two participants simultaneously. With the permission of the interviewees, the interviews were recorded and manually transcribed. All interviewees were given the opportunity to make corrections to the transcriptions, which did not lead to any major revisions.

3.2. Data analysis method

Systems theory and system safety concepts, as described in Section 2 of this paper, are used as the conceptual framework for our analysis. Through our analysis, we describe how observations in the interview data can be linked to this conceptual framework. This line of reasoning is called abductive reasoning. Abductive reasoning is a form of logical inference that involves making hypotheses to explain observed phenomena or data (Sætre & Van de Ven, 2021). In this paper, we hypothesize that ML systems can be best described through a systems

Table 1
Overview of interview IDs, interviewee positions, and organization types. The interviewees marked with an asterisk were employed by the same organization.

ID	Interviewee position(s)	Organization type
I1	Organizational manager	Executive organization applying ML
I2*	Data scientist	Executive organization applying ML
I3	Data scientist	Executive organization applying ML
I4*	Operational manager	Executive organization applying ML
I5	Data scientist	Executive organization applying ML
I6	Data scientist	Executive organization applying ML
I7	Operational manager	Agency supervising ML applications
I8	Organizational manager	Agency supervising ML applications
I9	Researcher	Agency supervising ML applications
I10	Operational manager	Agency supervising ML applications
I11	Operational manager and Advisor	Agency supervising ML applications

theory perspective, where we see an ML system as a complex system whose behavior can only be described by analyzing the system integrally instead of on a component level.

Through the analysis of the interview data, we find a set of perceptions about ML systems in the public sector. Following abductive reasoning, we expect that all perceptions will fit the theory. Furthermore, we may see some theory that is not supported by the perceptions of public professionals. Our hypothesis is wrong when we find observations that do not support, or contradict, the theory. Thus, we only expect to find observations that support the theory. Any observations that contradict or do not support the theory will be at odds with our hypothesis.

We analyzed the interview data using the software ATLAS.ti (version 23). An initial set of codes was directly derived from the conceptual framework presented in Section 3. In its entirety, the process of coding was done in five steps, based on the coding practices presented in Boyatzis (1998) and Braun and Clarke (2006):

1. Create an initial theory-based codebook. This initial codebook is based on the key system safety concepts which are introduced in Section 2. Five code groups were established. The first two groups, i.e. 'institutional factors' and 'technical factors', relate to the risk factors that follow from the institutional context of an ML system and the ML system itself. The third code group, 'mitigation strategies', relates to the concepts of safety management and control strategies. The fourth code group, 'control conditions', relates to the four control conditions that need to be met to remain in a safe system state. The last code group, 'accidents', relates to the situations in which the working of the system leads to an unwanted and unplanned event resulting in a loss.
2. Review and revise the codebook. A first round of coding was performed, going through all transcriptions while assigning codes to relevant pieces of text. During this first round, 18 codes were added to capture interesting results that relate to the risks of ML systems, resulting in a codebook with a total of 41 codes with a total of 128 quotations.
3. Check the reliability of the coder and codes. After the first round of coding, a second coder performed a round of coding, resulting in 51 codes. All documents were analyzed again with this set of codes, resulting in 328 quotations.
4. Validate codes and quotations. For codes with few quotations, it was checked whether they could be merged with other codes. Furthermore, quotations were revisited to check whether assigned codes were still valid for the text, building upon the experiences and insights from steps one to four. The resulting codebook can be found in Supplement 2.
5. Identify themes. Lastly, codes are grouped into themes. This process of thematic analysis can summarize the key features of the interview data and is useful to identify similarities and differences amongst the observed perceptions of the interviewees (Braun & Clarke, 2006). The identified themes are reported as the categories of risk factors in Section 4 of this paper.

Table 2 shows three examples of quotes with their related system safety concepts and assigned codes. There is much overlap between codes and concepts. It was very common to discuss control strategies and risk factors simultaneously. The lack of a control strategy can be interpreted as a risk factor, as seen in quote #2. Similarly, remarks about control conditions are often found in connection with risk factors, as seen in quote #3.

The coding allows us to systematically report on public professionals' perceptions of ML systems in the public sector. Since our starting point for coding follows directly from the concepts of system safety, we are subsequently able to compare the perceptions with the literature on these concepts. Through this comparison we can identify the parts of systems theory and system safety concepts without support of the observed perceptions. This is the main input for our discussion in Section 5.

Table 2
Examples of code assignments.

#	Quote	system safety concepts	Assigned codes
1	"In my opinion, there are few people who have the overall picture and can fully comprehend everything from A to Z, including all possible side effects that we may not want or that we did not expect." [I11]	Institutional factors Technical factors Control conditions	Lack of Knowledge System complexity Model condition
2	"It really depends on the culture within an organization. Does the organization truly want to improve, or is it more focused on checking off boxes? To what extent is the organization a learning organization or not?" [I7]	Institutional factors Control strategies	Lack of safety culture Lack of learning organization Safety culture Learning organization
3	"This means that someone who works with AI must know what the AI is doing, when it is reliable and when it is not, but also must be able to assess whether something is correct." [I10]	Control conditions Institutional factors	Model condition Observability condition Lack of knowledge

4. Results

In this section, we present the main findings from the interview analysis. We used the system safety concepts introduced in Section 2 of this paper as a starting point for coding. Through the identification of themes amongst the codes, five categories of risk factors are found that can lead to unsafe system states. Fig. 2 schematically depicts these categories within the conceptual framework as presented in Section 2.2. The risk factors are determined by the social and institutional context of the ML system and the ML system itself. The interviewees perceived both institutional and technical risk factors, which impact the control of the system and may lead to not meeting the control conditions, resulting in a hazardous system state with a probability of an accident occurring.

We use the following naming to refer to the categories of risk factors: organizational complexity, underdeveloped safety culture, a lack of knowledge, poor data and algorithm quality, and system complexity. Note that we implicitly define these categories by listing the observations that adhere to these categories. The following sections will elaborate on each of these factors and their consequences. Each section starts with the observations from the interviews, followed by a reflection on these results using system safety literature. For the latter, we identify relevant system safety concepts, which are marked in italics. Furthermore, in Section 4.6, we show how the interviewees perceive how these risk factors impact the ability to control the ML system and what accidents and corresponding losses are expected when control fails. We end the section with an overview of these categories, the perceived risk factors, and the identified relevant system safety concepts.

4.1. Organizational complexity

This section elaborates on the insights provided by the interviewees related to the institutional risk factor of 'organizational complexity.' Interviewees see organizational complexity as a factor leading to risks in working with ML systems. The following paragraphs show how this organizational complexity is perceived to materialize, after which we end this section by explaining organizational complexity through system safety concepts.

4.1.1. Involvement of multiple stakeholders

Several interviewees mentioned that it is key to let multiple stakeholders have a say in the development and use of ML systems [I2, I5, I6,

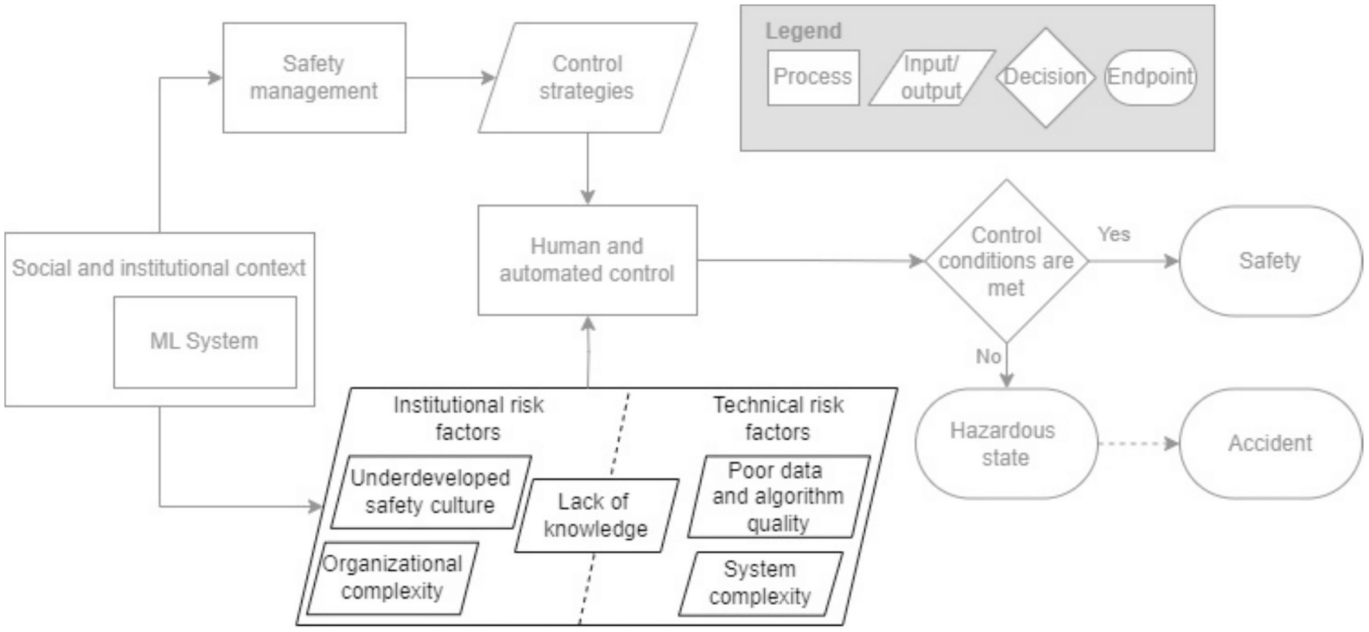

Fig. 2. Schematic representation of the perceived risk factors as part of the conceptual framework presented in Fig. 1.

I9, I11]. Civil servants must work in accordance with political decision-making but simultaneously have to adhere to the wishes of citizens that are not captured in laws and regulations [I9, I11]. One of the interviewees stated that *"the biggest risks lie in not taking into account the interest of individual citizens in the design* [of ML systems] *and in the way you can react* [to decisions following from ML systems]*"* [I9]. Besides citizens, the civil servants working with the output of ML systems need to be included during the design process [I3, I5, I6]. On the one hand, this is important for the designers of ML systems to understand the impact of the system on daily operations [I3, I6]. On the other hand, involving these end-users is key to getting these civil servants to accept and trust ML systems [I5].

4.1.2. Delegation of public tasks over different organizations

ML systems may use data that is generated or registered by multiple organizations. This may lead to misconceptions about the meaning of certain data [I11]. Interviewees in interview I11 mentioned a case where a citizen constantly suffered from wrong decisions, and *"when we had everybody at the same table, we found out that the data was interpreted differently by the receiving party than intended by the issuing party."* This shows that the delegation of tasks over different organizations may lead to accidents when data is shared without proper metadata.

4.1.3. Shared responsibilities with private parties

The interviewees see that organizational complexity increases when private parties are involved in designing the ML system [I7, I10, I11]. Responsibilities are shared with this private party, making it harder to know who is accountable when something goes wrong with the system [I11]. Furthermore, these third parties are not as familiar with the context of the public organization for which the ML system is designed as the public servants themselves, which can lead to misunderstandings about the (intended) use of the ML system [I7].

4.1.4. Values changing over time

Interviewees I1, I2, I3, and I5 acknowledge that organizational values change over time and that this impacts how trade-offs are made for the design of ML systems. They refer to three trade-offs for ML system design. First, the trade-off between model performance and explainability is mentioned by multiple interviewees [I1, I3]. Both the ability to make a good decision, as well as explaining how this decision was generated are important to civil servants. However, interviewees observe that while more complex algorithms can generally produce more accurate outcomes, this complexity might lead to fewer insights into how these outcomes were generated. There are no norms or guidelines for when model performance and explainability are 'good enough.' According to interviewee I5, *"Everybody thinks it is very important to have transparent and explainable AI, but it is unclear what kind of explainability people require."* A second trade-off exists between transparency and confidentiality [I3, I5]. Some organizations are, to a certain extent, not able to share how they do their job because this might allow for malicious activities. To quote one interviewee: *"If people spontaneously start to comply, that would be great, but you don't want them to avoid us."* For example, organizations that have the task of finding cases of fraud are having difficulties in being open about their use of ML systems, as citizens can alter their behavior if they know what variables are used to detect fraud. Lastly, a trade-off between privacy and non-discrimination is observed. Interviewee I2 stated, as an example, that *"you are not able to test whether an algorithm is sexist if you don't include gender data."*

4.1.5. Responsibility attribution

The issue of responsibility attribution within organizations is mentioned as a factor that can lead to hazardous situations [I1, I2, I3, I7, I9, I10]. Several interviewees agree that only humans can be attributed with the responsibility for decisions that follow from working with ML algorithms. However, the normal line of hierarchy within the public sector might not suffice. Interviewee I1 mentioned that *"at the moment, there is no minister with an affinity for data science while departments are working with it. He/she is still ministerially responsible"*. Another interviewee mentioned that *"the one* [that is responsible] *should have the chance to prevent* [accidents]*"* [I3]. Besides the lack of affinity and the lack of room for intervention, interviewee I2 mentioned the *"artificial separation between IT and other matters"* as a responsibility issue. Responsibility for the safety of ML algorithms cannot be attributed to the technology alone and requires a role for the users and other organizational entities involved in the development, use, or governance of such systems. On the other hand, interviewee I7 mentioned overlapping responsibilities as a potential hazard. People who know that others are also responsible are more likely to think that *"it will probably be all right,"* while others might have the same idea.

4.1.6. Organizational complexity in system safety

When we link these results to system safety literature, we see an important notion about the impact of organizational complexity can be found in the concept of *control coordination*. Control coordination is important for safety, mostly in those areas where responsibilities overlap and at the boundary of a responsibility (Leplat, 1984). Thus, control becomes more challenging in the organizational situations described by the interviewees, where different organizations or teams within organizations each have a role in one process. *Responsibility overlap* may lead to conflicting decisions or advice or to controllers waiting on each other to take responsibility for a task (Leveson, 2011). Fundamental to translating safety requirements to functional operationalization is the *safety control structure*, which comprehensively describes how different processes and actors are involved in implementing, supervising, and maintaining vital safety control mechanisms (Leveson, 2011). In the lower levels of the safety control structure, the operational process is described, in which the ML system plays a vital role, and in which an idea is required of how every possible hazard may emerge in real time and what should be done to prevent it or mitigate its associated risks. Beyond the operational process, we can identify higher levels of the safety control structure, which have to do with maintenance, (re)design of the system and process, operational management, and beyond that, organizational management, supervision and law, policy and democratic/political deliberation and decision-making (Leveson, 2011). System safety approaches can also be used to identify hazards and causal scenarios on these organizational levels and components. Though sometimes less emphasized, a *gap analysis* may be vital to identify holes in the current design of the existing organizational and social safety control structure (Leveson, 2011, p. 232). Such gaps may then be translated to risks of varying priorities with associated policy recommendations or mitigation strategies.

4.2. Underdeveloped safety culture

This section discusses the perceptions of interviewees related to the second identified institutional factor: safety culture. All interviewees named the cultural features of an organization as a factor influencing the safety of ML systems. The following paragraphs show how these cultural features are perceived to materialize, after which we end this section by explaining safety culture through system safety concepts.

4.2.1. Learning from mistakes

Professionals need to be able to speak up and act when something goes wrong while using an ML system [I9, I11]. They should be critical and *"be firm to ask critical questions about the system"* [I11]. This may not be the case when professionals do not have enough knowledge to be critical, but also if the professional does not feel that there is room in the organization to be critical about its own processes and behaviors. Furthermore, it can be difficult to stay critical when there seems no need to do so: *"It has been working for years, so why should I criticize?"* [I6]. Being able to criticize is linked to an organizational culture where people are open to learning from mistakes [I5, I7]. Being fearful of making mistakes can lead to hazardous situations. It is preventing organizations from letting external parties reflect on their work. According to interviewee I7: *"If you want to innovate, you need to be a learning organization, and part of that is making mistakes. To know if you make mistakes, you need reflection"*. Supervisors see resistance from organizations to be open about their work as *"being checked is never fun"* [I7]. One interviewee sees that the requirement to be open leads to organizations *"taking all possible measures to not have to comply"* [I1]. The biggest driver for not wanting to be open seems to come from fear of public opinion and any political repercussions. Interviewee I5 referred to a case where a mistake led to a heavy political reaction, which still challenges the transparent development of ML algorithms four years after this mistake.

4.2.2. The use of checklists for compliance

Several interviewees identified the use of checklists as a potential source for hazardous situations [I5, I6, I7, I11]. Checklists can contribute to a culture where it is more important to tick every box instead of becoming aware of the actual risk factors involved in using a certain ML algorithm. Due to checklists, *"people stop thinking"* about safety [I7].

4.2.3. Commitment through investments

The lack of commitment of organizations to implementing ML algorithms safely is also mentioned as a source of potential hazards [I1, I2]. *"Doing this carefully needs a very big investment"*, according to interviewee I1. Committing and investing in ML algorithms is, on the one hand, needed to make safe systems, but on the other hand, precarious in the public sector. Numerous cases of governmental overspending on IT projects can be found, causing "*the public sector to be wary of large IT projects"* [I1]. This causes organizations to *"not get past a hobbyist level"* of working with ML algorithms [I1]. Investments are needed in the *"automation* [infrastructure]*, knowledge and management system"* [I1]. However, interviewees see that the government expects immediate returns on investments and expects to spend less money on processes where ML algorithms are implemented [I2, I4].

4.2.4. Safety culture and system safety

Three *layers of organizational culture* can be distinguished: its artifacts, its beliefs and values, and its underlying assumptions (Shein, 2004). All three layers should be addressed to ensure safety. Leveson (2011) explicitly warns of a *paperwork culture* where explaining the safety of a system on paper is pursued at the cost of the actual safety in the real world. This aligns with the perception of the interviewees, who criticize the ever-growing number of checklists that are available for the design of safe ML systems. Instead of striving for real safety, the goal is to check all the boxes.

Multiple interviewees highlighted the negative effects that the *blame culture* has on safety. The fear of being blamed for mistakes comes at the cost of organizations covering up their mistakes, inhibiting them from learning from these mistakes. Dekker (2012) describes this as the trade-off between accountability and safety. Although working towards a balance between these values is challenging, he describes steps that can be taken to move towards a *just culture*. System safety states that safety is ensured at the system level, considering all actors directly or indirectly involved (Leveson, 2011). This directly implies that when one wants to improve the safety of a system by learning from an accident, one should address the system as a whole instead of individual people. Blaming thus is not only standing in the way of being open and learning from mistakes, it also wrongfully leads to the idea that, when the blamed changes his behavior, safety will be ensured.

4.3. Lack of knowledge

This section discusses the perceived risk factor 'lack of knowledge,' which can be seen as an institutional risk factor, as well as a technical one, as the knowledge needed is partly directly related to the technical artifact of an ML system. Knowledge is a central and repeating factor throughout the interviews. The following paragraphs show how this lack of knowledge is perceived to materialize, after which we end this section by explaining the importance of knowledge through system safety concepts.

4.3.1. Lack of knowledge at the end-user

Knowing what an ML system does, what it is capable of, and what its limits are, are essential for an ML system's safe operation. According to interviewee I11, *"not [having] enough people that have the overview and understand the system from A to Z"* makes it impossible to *"have an overview of possible side effects"* of an ML system. Multiple times, the lack of knowledge leading to hazardous situations is attributed to the end-

users, the civil servants working with the outcomes of an ML system [I3, I5, I7, I8, I10, I11]. These end-users are tasked to work with ML systems' outcomes and serve as a 'human-in-the-loop,' responsible for *"checking whether it* [the output of an ML system] *is correct"* [I11]. However, interviewees indicate that a civil servant who is not trained in working with these systems will not perform meaningful checks and balances [I1, I8, I10, I11].

4.3.2. Deskilling through ML systems use

Applying ML within an organization can lead to decreased knowledge about the process for which an ML system is used. Seemingly correctly functioning ML systems can move an organization to hire *"cheap labor instead of a more expensive employee to check an AI system"* [I10]. This means that the skills to manually perform the tasks that are now carried out by the AI system will disappear.

4.3.3. Lack of knowledge at the managerial levels

A lack of knowledge about ML systems is also attributed to professionals in managerial roles [I1, I3, I5, I6, I8, I11]. One interviewee mentioned that *"the risk lies in the gap between the knowledge of analysts and decision-makers"* [I1]. It is stated that managers push for the use of ML systems because it is *"new and hip"* and so they can *"brag about who spends most on AI,"* while ML *"is not always the solution"* [I3].

4.3.4. Reasons for a lack of knowledge

Various reasons are mentioned for this lack of knowledge. One reason is that much knowledge about ML systems is gained from temporarily hiring external consultants and disappears when these consultants finish their temporary jobs [I2]. Another reason is the reluctance to share experiences of working with ML between organizations. According to interviewee I5: *"When you try hard to be transparent […] you will be punished most"*. This leads to fear of being open about using ML systems, which stands in the way of sharing experiences [I5, I7].

4.3.5. Lack of knowledge and system safety

A lack of knowledge inhibits a good overview of the potential risks of a system, which is key for establishing safety constraints. A lack of knowledge about the system especially creates difficulties for operators in dealing with non-routine events (Leveson, 2011). Furthermore, "human skill levels and required knowledge almost always go up" when supervising an automated system (Leveson, 2011, p. 229), such as an ML system. This can be explained by the fact that *human and automated control* can only be meaningful when the controller understands how the system works, which comes on top of the knowledge needed about the subject matter for which the system is being used. This notion of requirements for enhanced skill and expertise may conflict with the initial reason to start using ML systems, which is often the promise of reduced operational costs.

It is vital to distinguish the reasons why a lack of knowledge arises that contributes to safety hazards and accidents. As we saw in the interviews, often a lack of knowledge is projected onto the operator or user of a particular system – those operating the process in which an ML system is used. However, such knowledge deficiencies can only be meaningfully attributed if there is a clear understanding of what knowledge was needed to safely and adequately operate the process and ML system in the first place. The history of system safety shows that most often, *operator error* is not a function of the operator's capabilities but of the environment in which they are asked to operate (Leveson, 2011). The environment determines what information is available to the operator and at what time; it determines what actions and support are available. Moreover, in the case of a complex system that requires nuanced knowledge to understand and operate, those responsible for managing and developing the system should have a clear idea of what knowledge and capabilities are required to operate it safely.

4.4. Poor data and algorithm quality

There were different opinions amongst the interviewees about whether data and algorithm quality are relevant factors impacting the safety of ML algorithms. Interviewee I10 stated: *"I don't think that it often goes wrong in the technical details."* However, most interviewees identified two potential risk factors related to the quality of data and algorithms. In this section, we will first address these two risk factors, 'bias' and 'model performance,' after which we end this section by explaining these risk factors through system safety concepts.

4.4.1. Bias

The interviewees describe several forms of bias as a problem that can lead to hazards [I1, I3, I5, I6, I8, I11]. Fraud detection is mentioned as a field with a lot of potential for bias [I1, I5, I6]. ML algorithms may be able to detect fraud for a specific subset of cases but may fail to detect fraud in others. Retraining the ML might then propagate the bias towards this specific subset [I5]. Furthermore, interviewees mention the possibility that historical data contains bias due to the way data was registered [I8] or the bias that human operators had when a task was carried out without the ML algorithm [I1, I3, I5]. Confirmation bias can be introduced or maintained by human operators working with ML algorithm advice because *"they only follow the model when they agree, and in this way strengthen their own feelings"* [I3].

4.4.2. Model performance

Model performance may be a source of hazardous situations [I1, I2, I4, I7, I10]. Interviewees describe cases where important variables were not included in the ML algorithm, leading to bad predictions [I2, I4]. Furthermore, limited data quality is mentioned as a reason for underperforming ML systems [I2, I10]. Low model performance may be a reason for ML systems not being put into production, which would not lead to accidents. However, one interviewee described a case where an ML algorithm was used, although accuracy metrics were very low [I7].

4.4.3. Data and algorithm quality and system safety

Although accidents are not caused solely by technical components of ML (Dobbe, 2022), such as data and algorithms, they can impact the safety of a system. However, the notion of bias presumes the *programmability* of the 'correct' decision. In many situations, we know additional circumstances are needed to properly understand what a safe and just decision is (Dobbe et al., 2018). Furthermore, a crucial understanding of system safety is that *reliability* or accuracy is not sufficient and may not be necessary for safe outcomes (Leveson, 2011).

To first address the latter notion of necessity, it is more important to understand how issues of data or algorithm quality may lead to hazardous situations and harmful outcomes. Put differently, it is unwise to assume one can prevent errors in data or algorithmic outputs (Gansky & McDonald, 2022). Instead, it is crucial to assume such errors are made and to have control mechanisms in place to ensure that when an ML system fails, it *fails safely* or is prevented from being used in a consequential manner (Leveson, 2011). Second, and more pertinent to safety, while improving data and algorithm quality is wise, a sole emphasis on quality is insufficient to ensure safe outcomes, as the other emergent risk factors listed in this section have shown.

4.5. System complexity

Systems using ML algorithms tend to be complex in a way that imposes risks for the safety of a system. Interviewees distinguish complexity on three different levels. In this section, we will first address these levels, after which we end this section by explaining system complexity through system safety concepts.

4.5.1. Complexity at the process level

System complexity is found at the process level in which one or

multiple ML algorithms can be used [I6, I7, I8, I9, I10, I11]. *"I don't think we can comprehend how huge the processes and used data are,"* according to interviewee I11. Stacking regulations makes public tasks more and more complex, which makes it very hard for civil servants to fully oversee these tasks [I11]. The use of ML algorithms that use multiple sources of data can make the system even more complex. More complexity is introduced when third parties deliver the ML algorithm [I6].

4.5.2. Complexity at the algorithm level

The ML algorithm itself can become very complex [I1, I3, I5, I6, I7, I9, I11]. ML algorithms, such as neural networks, have complex internal structures that are difficult for humans to understand [I5]. This opacity is seen as an obstruction for civil servants to check whether outputs make sense, which is needed for these professionals in order to be critical of ML outputs [I11]. Besides this, governmental organizations often have the responsibility to *"go back in time and see how a specific decision came to be"* [I9]. There are methods to explain model output. However, *"it is very hard to determine what is a good explanation"* [I5].

4.5.3. Complexity at the data level

System complexity can be found in the data that is used in ML algorithms [I5, I10]. Some data is *"very specialistic,"* making it hard even for data scientists to understand what it means: *"I understand only part of that data"* [I5]. Some ML algorithms use data that is the output of other ML algorithms. This introduces the hazard that if one ML algorithm outputs a bad result, the second ML algorithm will take this bad result as an input, possibly causing more bad outputs [I10].

4.5.4. System complexity and system safety

The increase in system complexity is one of the reasons why a new approach towards safety engineering was required (Leveson, 2011, p. 3). As system complexity grows, the probability increases that accidents occur as a result of the interaction between system components. This asks for an approach towards safety that goes beyond safety on the component level and addresses *system level safety*, which system safety provides.

Software, in general, suffers from the *curse of flexibility* (Leveson, 2011). The physical restraints of narrow-purpose machines have been lifted with the introduction of the computer and the increasing ease with which software can be developed. This makes it very easy to create complex systems, for which it becomes increasingly difficult to implement and test safety requirements. In these complex systems, it becomes impossible to foresee all system states, including the ones that may lead to accidents. The recent experimentation with large language models, which are typically comprised of trillions of parameters or more and which are inscrutable due to their complexity (and potentially also due to these being hidden behind corporate APIs or interfaces), provides a case-in-point (Dobbe, 2022). Often, ML system projects suffer a tendency to start coding without properly understanding the broader requirements and how an ML system will be situated in context. Data experts tend to lack the *needed domain expertise* to design the system in its broader context (Leveson & Weiss, 2009). Once safety issues or requirements arise, it can then be costly to adjust the system design.

4.6. Unsafe control, accidents, and loss

The risk factors described in Sections 4.1 to 4.5 can lead to hazardous situations or system states of the ML algorithm. Separately or in combination, these factors can hinder the safe control of the system. In other words, one or more control conditions, as described in Section 2.2, cannot be met due to one of the risk factors. For example, the 'action condition' may not be met when a culture is lacking in which civil servants can criticize the system and procedures that they have to work with [I11]. The 'model condition' may not be met when operators have little knowledge about how the ML system works [I3, I5, I7, I8, I10, I11]. The 'observability condition' may not be met when operators are faced with high system complexity, on the level of the technical artifact [I1, I3, I5, I6, I7, I9, I11], the data [I5, I10] or on the process level [I6, I7, I8, I9, I10, I11].

Under unsafe control, a system enters a hazardous state, which has a probability of leading to an accident. The interviewees describe discriminatory decisions as one of these accidents [I3, I4, I6, I8, I11]. When a certain group of citizens is checked more often, or decisions are incorrect for a specific group, this is seen as an accident. Furthermore, losing room for considering individuals' personal circumstances is described as an accident [I11], as well as privacy breaches [I1, I2, I5, I10]. Interviewees describe a loss of trust in governmental organizations due to these accidents [I9, I10, I11].

4.7. Overview of results

Sections 4.1 to 4.5 present the results of the analysis of the interview data. Furthermore, we reflected on the perceptions of the interviewees by linking these perceptions to relevant system safety concepts. Table 3 presents these results in summary, listing the risk factor categories, risk factors, and the system safety concepts to which we linked these risk factors.

5. Discussion

In the previous section, we show that the public professionals who are interviewed perceive organizational complexity, an underdeveloped safety culture, a lack of knowledge about ML systems, poor data and algorithm quality, and system complexity as risk factors that lead to unsafe control and potential accidents. These perceptions are linked to concepts and lessons from system safety literature. In this section, we identify four constitutive factors of system safety that are currently underemphasized or missing when the perceptions of the interviewees are compared with the system safety literature. We discuss these gaps in Section 5.1. Furthermore, in Section 5.2, we link back to our initial hypothesis, being that ML systems are best described through a systems theory perspective.

Table 3
Overview of the perceived risk factors per risk factor category, and the identified relevant system safety concepts.

Risk factor category	Risk factor	Relevant system safety concepts
Organizational Complexity (Section 4.1)	Involvement of multiple stakeholders Delegation of public tasks over different organizations; Shared responsibilities with private parties; Values changing over time; Responsibility attribution.	• Control coordination • Responsibility overlap • Safety control structure • Gap analysis
Underdeveloped safety culture (Section 4.2)	Learning from mistakes The use of checklists for compliance Commitment through investments	• Layers of organizational culture • Paperwork culture • Blame/just culture
Lack of knowledge (Section 4.3	Lack of knowledge at the end-user; Deskilling through ML systems use; Lack of knowledge at the managerial levels;	• Human and automated control • Operator error
Poor data and algorithm quality (Section 4.4	Bias Model performance	• Programmability • Reliability vs. safety • Fail safely
System Complexity (Section 4.5)	Complexity at the process level Complexity at the algorithm level Complexity at the data level	• System level safety • Curse of flexibility • Need for domain expertise

5.1. Gaps between perceptions and system safety

We identify four constitutive factors of system safety that are not addressed by the interviewed public professionals. We argue that these factors are key when designing and operating safe ML systems.

5.1.1. Explicit safety goals and constraints

To design and operate a safe system, it should be clear what accidents need to be prevented. This is needed as input for any trade-off that is made between the safety of the system and any other potential goal, such as increased efficiency or the reduction of costs. Not being clear about the level of safety that is expected will inhibit the first control condition as formulated in Section 2.2, as the controller needs to be aware of the goal of his/her control actions and what outcomes to prevent.

Preventing unsafe behavior in a system can be done by implementing safety constraints. In engineering, these constraints often have a physical nature. For example, a passenger train is not allowed to move when a door is open (Leveson, 2011, p. 192). This constrains the operation of the train with the aim of preventing hazardous situations in which passengers can exit the train during the ride. Constraints can also be designed for ML systems in the public sector if it is clear what hazardous situation we wish to avoid.

The emergent nature of ML systems means that certain hazards may not have been anticipated and present themselves over time. This warrants explicit procedures, in addition to safety hazard analysis, to identify and follow up on issues during operation. Establishing such procedures in democratic contexts requires adequate avenues for expressing dissent (Dobbe et al., 2021) and ensuring follow-up to build and maintain trust (Leveson, 2011).

5.1.2. Inclusion of systemic factors in system design

The data in this paper shows that public professionals acknowledge non-technical factors when thinking about ML system safety. However, this acknowledgment has not yet led to structurally including these factors in design and governance practices. Interviewees do identify biased data as a potential source for hazards but address this as a problem that has to be solved within the technical artifacts of the ML system. Rather, biases existed in data before this data was used for training ML algorithms (Dobbe et al., 2018). Focusing on de-biasing ML algorithm output is therefore a suboptimal path for countering bias and can better be addressed where the data finds its origin.

The influence and importance of systemic factors have recently been highlighted in research (e.g., Rodriguez Rivas-Stellaard, 2023) and parliamentary investigations (e.g., Tweede Kamer der Staten-Generaal, 2023) into several Dutch policy scandals. Although these scandals do not necessarily revolve around the use of ML algorithms, they show how systems, scoped from politics and policymaking towards the real impact of public decision-making on individual citizens, tend to fail. Evaluations of failed policy "tend to focus on pressing issues at hand in the here and now" (Rodriguez Rivas-Stellaard, 2023) and subsequently provide patchwork solutions for previous policy problems instead of fixing systemic problems.

5.1.3. Development of safety control structures

Although the interviewees recognize the importance of clear responsibilities for controlling an ML system, we see that the development of safety control structures is underemphasized in the interviews and in the ongoing discourse on ML system safety. Systems theory describes the control of systems as hierarchically imposed constraints from one level on the activity on the level below (Checkland, 1981). Leveson (2011, p. 82) shows that for sociotechnical systems, this control structure comprises several layers of control, including, from higher to lower levels: legislators, regulators, company management, project and operations management, and human and automated controllers in the operational process of the system, each imposing their constraints on the layer below.

Although these levels of sociotechnical control can be found in European governance structures, it is often not formalized what constraints are imposed from layer to layer. Regarding the control structure of ML systems in the Netherlands, the Dutch Data Protection Authority was instated as the 'algorithm watchdog' by parliament. This was, however without extending the mandate of the authority, which is legally still bound by the GDPR. Meanwhile, several auditors and supervisors have published reports about the governmental use of ML systems, but it remains unclear how responsibilities are divided amongst them.

Interviewees mentioned that citizens have to be informed about the way the government is using ML systems and described citizens as controllers. However, to maintain meaningful control, the four control conditions described in Section 2.2 must be met, which is currently unrealistic. Citizens do not have insight into the risks of ML systems, do not understand how these systems work, have insufficient means to prevent or take action, and systems often do not work for them.

5.1.4. Migration to higher risk

During their operation, systems *"tend to migrate toward states of higher risk"* (Leveson, 2011, p.52; Rasmussen, 1997). Although the technical part of ML systems is generally perceived as static artifacts (i.e., the computer code does not change), this cannot be said about the sociotechnical system and the context of the system. External pressure from, for example, politics can lead to a focus shift from safety towards cost reduction or efficiency gains. Furthermore, it is easy to forget why safety constraints are in place when accidents do not happen, which can motivate overstepping or ignoring established constraints and safety control mechanisms.

Ensuring safety means that migration towards states of higher risk should be accounted for. First, this can be done during the design of a system, as this migration can be expected. Controls can be implemented that limit the possibility of this migration. Second, migration can be detected during the operation of a system. This requires feedback loops in which signals of system behavior that diverts from the original system design are communicated, after which evaluation of the safety constraints on this new behavior is needed.

Most often, such migration to higher states of risk is of social or political nature and is first expressed in the behaviors and decisions made at an organizational level. This tendency was starkly observed in the context of the Boeing 737-MAX crashes, which occurred in 2018–2019. While safety management, culture, and oversight are central pillars to ensuring safety in aviation, investigations by the House Committee on Transportation and Infrastructure point out that migration to unacceptable risk levels occurred: "The MAX crashes were not the result of a singular failure, technical mistake, or mismanaged event. They were the horrific culmination of a series of faulty technical assumptions by Boeing's engineers, a lack of transparency on the part of Boeing's management, and grossly insufficient oversight by the FAA—the pernicious result of regulatory capture on the part of the FAA with respect to its responsibilities to perform robust oversight of Boeing and to ensure the safety of the flying public" (The House Committee on Transportation and Infrastructure, 2020, p. 6). This case-in-point underlines the importance of monitoring the behavior of key actors responsible for a system's safety, including those responsible for oversight.

5.2. Applying systems theory to ML systems

This paper adopted the hypothesis that ML systems generally behave in a way that is described by systems theory. This hypothesis is based on the observation that the behavior of ML systems can only be explained by analyzing its components in cohesion. Our analysis of interview data shows that the perceptions of public professionals align with our hypothesis. Public professionals confirm that the context in which the technical artifacts of ML systems are deployed highly influences the risks

that we can expect from this system. They see the complexity of the ML system, both in the technical and the organizational components of the system, as an important risk factor. Omitting this complexity and simplifying ML systems by assessing its components individually thus leaves out key characteristics of the system's emergent behavior.

6. Limitations and implications

This paper presents a first exploration of the use of systems theory and system safety concepts for assessing the risks of ML systems in the public sector. The method and the data used in this study present limitations, which we will address in this section. Furthermore, in this section we address the implications of our research for both science and policy.

6.1. Limitations and recommendations for future research

The nature of this study presents limitations for interpretation. The exploratory character of the study limits us to making statements about the completeness of our data. Although we argue that our sample is a representative sample of Dutch public professionals working either with ML systems or being tasked with supervising the use of such systems, new observations may arise when other professionals are interviewed. Future research may further strengthen and broaden the empirical evidence on risk factors of ML systems in the public sector. Furthermore, our data collection was focussed on the public sector. System safety, however, is not bound to the public sector. Using system safety concepts to adress ML risk factors may be usefull for addressing challenges in the private sector in future research.

A limitation of our research is the geographic and temporal specificity of our data. Contexts, other than the Dutch context in the current stage of ML system development, may present nuances and contextual factors relevant to the safety of Machine Learning systems that are unique to other countries or regions, which our study does not capture. The insights gained from Dutch professionals are highly valuable, given the country's proactive stance on ML adoption and regulation, similar to other northern European countries. However, it is important to recognize that the findings might differ in contexts where governmental approaches to ML are at different stages of development or follow different regulatory and operational frameworks, such as the USA or Singapore. Future research should aim to include a more diverse range of geographic contexts to ensure a comprehensive understanding of the safety of ML systems in the public sector. Nonetheless, many of the systemic risk factors and safety considerations identified in our study are likely to be relevant across different contexts, given the universal nature of systems theory and system safety principles.

Our abductive line of reasoning shows that a systems theory lens is highly suitable when analyzing ML system safety. However, abductive reasoning does not give us a definite answer to whether systems theory is the best, or the only lens, to adopt for this analysis. We recommend that future research addresses this limitation by explicitly using and building upon existing theories, other than systems theory.

Although the system safety lens provides us with new insights related to ML system safety, its true powers lie at the level of individual systems. We believe that this presents several fruitful avenues for further research. Leveson (2011) describes methods for the design of safe systems as well as for the analyses of accidents. Testing these methods will show their applicability to ML systems, and the results may generate lessons for the design and operation of safe ML systems. Furthermore, dealing with the risk factors that we have identified requires capability building within organizations. This process may not be straightforward, as Dekker (2012) shows for the path towards a 'just culture'. The insights from research into the process of this capacity building may provide an extra layer of depth to the field of system safety. Lastly, there may be cultural nuances between different countries that may affect the trade-offs related to safety. These differences may include, for example, risk perception and blame culture. Studying these differences can provide insights that are key to safe ML systems, as culture is an important institutional factor that constitutes safety.

The exploration presented in this paper may serve as a starting point for more in-depth empirical research of the individual risk factors presented in Section 4. We will briefly highlight three examples of such research opportunities. First, public organizations are struggling to design processes for developing compliant ML systems. We observe that checklists are popular tools but can be at odds with a healthy safety culture. Testing and validating alternatives to checklists can provide insights for both academics and practitioners. Second, we see in our data a shared perception amongst public professionals that a lack of knowledge is a risk factor. Experiments have shown the importance of experience for ML system operators (Janssen et al., 2022), but more empirical research can make valuable impacts on ML safety. For example, case studies could point out how organizations train their operators and how this affects the quality of decision-making. Third, we see avenues for further developments regarding handling ML system mistakes. ML systems will make mistakes, so it would be wise to not only strive for prevention but also look at how to go about these mistakes. Both correcting the mistake and prevention of similar mistakes in the future are key here. Academic research may provide opportunities to (anonymously) share best practices, even though sharing information on the mistakes may be sensitive.

6.2. Contributions to science

This study builds on systems theory and system safety concepts that have had significant impact within a plethora of academic fields. Systems theory is, however, scarcely used explicitly to research information systems. Studies do implicitly use some of its assumptions, for example, by referring to information systems as 'socio-technical systems' (Janssen & Kuk, 2016; Kolkman, 2020). In this paper, we show the perils of harnessing systems theory and more specifically the concepts of system safety. We extensively use the work of Leveson (2011), whose work influences several engineering sectors such as the space, aviation, and automotive sector. Leveson and Weiss (2009) show the applicability of system safety for software, but a direct link with ML systems in the public sector is not made. In this paper, we build on this literature, and more recent conceptual work that links system safety to ML systems (e.g., Dobbe, 2022). Here, we followed the recommendation of Zuiderwijk et al. (2021) to perform more empirical studies.

Our study contributes to existing literature by introducing new empirical data and analyzing this data using concepts with a strong theoretical lens. This theoretical lens allows to disentangle different risk factors and show how they affect the safety of ML systems in the public sector. Furthermore, our analysis goes beyond the implementation phase of ML systems and explicitly includes the risks of ML systems when they are deployed in real life settings, which is key if we view safety as an emergent property of an ML system (Dobbe et al., 2018).

6.3. Implications for policy

The importance of including organizational and management components as an explicit part of safety analysis, design, and governance cannot be overstated. However, it often requires overcoming a tendency to attribute safety to technical artifacts and expertise alone. As a result, organizational components may be overlooked, both as contributing directly to safety or indirectly as a systemic factor of relevance. Therefore, we recommend that policymakers adopt a system lens when considering safety and designing and operating ML systems. This includes the design of control structures that transcend the operation of the technical artifact and that counterbalance pressure for efficiency improvements and cost reductions that are at odds with safety constraints.

The required knowledge for safely and effectively operating ML

systems in government processes extends beyond the operator and technical experts involved in building the associated software. Once the need to establish system safety standards and responsibilities is acknowledged, the rich history of system safety approaches may serve as a critical path to perform gap analyses in safety management. Consequently, the necessary knowledge across the organizational entities and roles involved in developing, implementing, using, managing, overseeing, and otherwise governing a process and associated ML systems can be identified. We recommend viewing the safety of ML systems as an organizational challenge rather than that of the data scientists. Accordingly, increasing the knowledge and expertise needed for the safe operation of ML systems should include a wide span of stakeholders within and beyond the organization.

System safety concepts can be used directly to assess policy initiatives to govern ML systems critically. We will give two examples of policy initiatives and their respective pitfalls that one can expect when adopting a system safety lens. First, the introduction of guidelines for ML systems. The context dependency of ML systems makes it difficult to introduce generic guidelines. This may be the reason for these guidelines to focus primarily on the technical artifacts and the design phase of ML systems. Through systems theory, we would expect that ML systems will show emergent behavior that cannot be predicted by assessing the technological artifacts in isolation and behavior should be continuously monitored in an inherently changing context. Second, the introduction of transparency requirements. Although transparency is regarded as a main ingredient for democratic processes (König & Wenzelburger, 2020), we see that transparency can lead to a backlash. This backlash can lead to a culture of fear for publicity and lack of willingness to share about ML practices. Instead, a culture in which learning from mistakes is promoted is key for safety. Although there is no silver bullet for this trade-off between transparency and safety, there are pathways towards a more 'just culture' (Dekker, 2012), and references can be taken from sectors with more maturity in dealing with this trade-off, such as the healthcare or transportation sector.

7. Conclusions

In this paper, we identified the overlap and gaps between the perceptions of public professionals and concepts of system safety regarding risk factors of ML systems in the public sector. We hypothesized that ML systems safety can be best addressed through a systems theory lens. Subsequently, we used systems theory and key concepts from the system safety literature to analyze new and existing interview data. From the interview data, we identified organizational complexity, an underdeveloped safety culture, a lack of knowledge, poor data and algorithm quality, and system complexity as constitutive risk factors that are present both in the perceptions of public professionals and can be traced back to system safety lessons. Furthermore, we identify that the need for explicit safety goals and safety constraints, safety control structures as well as the tendency for systems to migrate to a state of higher risk are missing in the perceptions of the interviewed professionals. We argue that these insights are key when designing policy instruments that aim to make ML systems in the public sector safe.

We show that a systems theory lens is fitted for addressing ML systems and system safety provides tools for increasing the safety of ML systems. Therefore, we see opportunities for further researching ML systems and their challenges for the public sector using systems theory and system safety concepts and recommend conducting more empirical studies regarding the risk factors presented in this paper. Policy makers may also benefit from adopting a systems theory lens when designing policies that aim for ML system safety. Our research implies that policies should address the organizational challenges of ML system safety and that system safety concepts can be used to critically assess policy initiatives.

CRediT authorship contribution statement

J. Delfos: Writing – original draft, Visualization, Validation, Software, Resources, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **A.M.G. Zuiderwijk:** Writing – review & editing, Validation, Supervision, Methodology. **S. van Cranenburgh:** Writing – review & editing, Validation, Supervision, Methodology. **C.G. Chorus:** Writing – review & editing, Supervision. **R. I.J. Dobbe:** Writing – review & editing, Validation, Supervision, Methodology, Formal analysis.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.giq.2024.101963.

References

Aoki, N. (2020). An experimental study of public trust in AI chatbots in the public sector. *Government Information Quarterly, 37*(4), Article 101490. https://doi.org/10.1016/j.giq.2020.101490

Artificial Intelligence Act (2024).

Boyatzis, R. E. (1998). *Transforming qualitative information: Thematic analysis and code development*. Sage Publications, Inc.

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology, 3*(2), 77–101. https://doi.org/10.1191/1478088706qp063oa

Brennan, T., Dieterich, W., & Ehret, B. (2009). Evaluating the predictive validity of the compas risk and needs assessment system. *Criminal Justice and Behavior, 36*(1), 21–40. https://doi.org/10.1177/0093854808326545

Broeders, D., Schrijvers, E., van der Sloot, B., van Brakel, R., de Hoog, J., & Hirsch Ballin, E. (2017). Big data and security policies: Towards a framework for regulating the phases of analytics and use of big data. *Computer Law and Security Review, 33*(3), 309–323. https://doi.org/10.1016/j.clsr.2017.03.002

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

Bryson, J., & Winfield, A. (2017). Standardizing ethical design for artificial intelligence and autonomous systems. *Computer, 50*(5), 116–119. https://doi.org/10.1109/MC.2017.154

Cath, C., & Jansen, F. (2022). Dutch comfort: The limits of AI governance through municipal registers. *Techné: Research in Philosophy and Technology, 26*(3), 395–412. https://doi.org/10.5840/techne202323172

Cath, C., Wachter, S., Mittelstadt, B., Taddeo, M., & Floridi, L. (2018). Artificial intelligence and the 'good society': The US, EU, and UK approach. *Science and Engineering Ethics, 24*(2), 505–528. https://doi.org/10.1007/s11948-017-9901-7

Checkland, P. (1981). *Systems thinking, systems practice*. John Wiley & Sons.

Dekker, S. (2012). *Just culture: Balancing safety and accountability* (2nd ed.). Ashgate Publishing, Ltd.

Desouza, K. C., Dawson, G. S., & Chenok, D. (2020). Designing, developing, and deploying artificial intelligence systems: Lessons from and for the public sector. *Business Horizons, 63*(2), 205–213. https://doi.org/10.1016/j.bushor.2019.11.004

Dobbe, R. (2022). System safety and artificial intelligence. In J. B. Bullock, Y.-C. Chen, J. Himmelreich, V. M. Hudson, A. Korinek, M. M. Young, & B. Zhang (Eds.), *The Oxford Handbook of AI Governance* (pp. 441–458). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780197579329.013.67.

Dobbe, R., Dean, S., Gilbert, T., & Kohli, N. (2018). A broader view on bias in automated decision-making: Reflecting on epistemology and dynamics. In *2018 workshop on fairness, accountability and transparency in machine learning during ICML 2018*. https://doi.org/10.48550/arXiv.1807.00553

Dobbe, R., Krendl Gilbert, T., & Mintz, Y. (2021). Hard choices in artificial intelligence. *Artificial Intelligence, 300*, Article 103555. https://doi.org/10.1016/j.artint.2021.103555

Dutch Data Protection Authority. (2023). Algoritmetoezicht. https://www.autoriteitpersoonsgegevens.nl/nl/onderwerpen/algoritmes/algoritmetoezicht.

Dwivedi, Y. K., Hughes, L., Ismagilova, E., Aarts, G., Coombs, C., Crick, T., … Williams, M. D. (2021). Artificial intelligence (AI): Multidisciplinary perspectives on emerging challenges, opportunities, and agenda for research, practice and policy. *International Journal of Information Management, 57*(July), Article 101994. https://doi.org/10.1016/j.ijinfomgt.2019.08.002

Engstrom, D. F., Ho, D. E., Sharkey, C. M., & Cuéllar, M.-F. (2020). Government by algorithm: Artificial intelligence in federal administrative agencies. *NYU School of Law, Public Law Research Paper, 20–54*. https://doi.org/10.2139/ssrn.3551505

Fountain, J. E. (2022). The moon, the ghetto and artificial intelligence: Reducing systemic racism in computational algorithms. *Government Information Quarterly, 39*(2), Article 101645. https://doi.org/10.1016/j.giq.2021.101645

Gansky, B., & McDonald, S. (2022). CounterFAccTual: How FAccT undermines its organizing principles. In *2022 ACM conference on fairness, accountability, and transparency* (pp. 1982–1992). https://doi.org/10.1145/3531146.3533241

Gorman, G. E., Clayton, P. R., Shep, S. J., & Clayton, A. (2005). *Qualitative research for the information professional: A practical handbook* (2nd ed.). Facet Publishing.

Haesevoets, T., Verschuere, B., Van Severen, R., & Roets, A. (2024). How do citizens perceive the use of artificial intelligence in public sector decisions? *Government Information Quarterly, 41*(1), Article 101906. https://doi.org/10.1016/j.giq.2023.101906

High-Level Expert Group on AI. (2019). Ethics guidelines for trustworthy AI. https://ec.europa.eu/futurium/en/ai-alliance-consultation.1.html.

Hoffmann, A. L. (2019). Where fairness fails: Data, algorithms, and the limits of antidiscrimination discourse. *Information, Communication & Society, 22*(7), 900–915. https://doi.org/10.1080/1369118X.2019.1573912

Janssen, M., Brous, P., Estevez, E., Barbosa, L. S., & Janowski, T. (2020). Data governance: Organizing data for trustworthy artificial intelligence. *Government Information Quarterly, 37*(3), Article 101493. https://doi.org/10.1016/j.giq.2020.101493

Janssen, M., Hartog, M., Matheus, R., Yi Ding, A., & Kuk, G. (2022). Will algorithms blind people? The effect of explainable AI and decision-Makers' experience on AI-supported decision-making in government. *Social Science Computer Review, 40*(2), 478–493. https://doi.org/10.1177/0894439320980118

Janssen, M., & Kuk, G. (2016). The challenges and limits of big data algorithms in technocratic governance. *Government Information Quarterly, 33*(3), 371–377. https://doi.org/10.1016/j.giq.2016.08.011

Kleizen, B., Van Dooren, W., Verhoest, K., & Tan, E. (2023). Do citizens trust trustworthy artificial intelligence? Experimental evidence on the limits of ethical AI measures in government. *Government Information Quarterly, 40*(4), Article 101834. https://doi.org/10.1016/j.giq.2023.101834

Kolkman, D. (2020). The usefulness of algorithmic models in policy making. *Government Information Quarterly, 37*(3), Article 101488. https://doi.org/10.1016/j.giq.2020.101488

König, P. D., & Wenzelburger, G. (2020). Opportunity for renewal or disruptive force? How artificial intelligence alters democratic politics. *Government Information Quarterly, 37*(3), Article 101489. https://doi.org/10.1016/j.giq.2020.101489

Koppenjan, J., & Groenewegen, J. (2005). Institutional design for complex technological systems. *International Journal of Technology, Policy and Management, 5*(3), 240. https://doi.org/10.1504/IJTPM.2005.008406

Kroll, J. A., Huey, J., Barocas, S., Felten, E. W., Reidenberg, J. R., Robinson, D. G., & Yu, H. (2015). Accountable algorithms. *University of Pennsylvania Law Review, 165*(633), 633–705. http://arks.princeton.edu/ark:/88435/dsp014b29b837r.

Leplat, J. (1984). Occupational accident research and systems approach. *Journal of Occupational Accidents, 6*(1–3), 77–89.

Leveson, N. (1995). *SafeWare : System safety and computers*. Addison-Wesley.

Leveson, N. (2011). In J. Moses, R. De Neufville, M. Heitor, G. Morgan, E. Paté-Cornell, W. Rouse, & Flexibility (Eds.), *Engineering a safer world - systems thinking applied to safety*. The MIT Press. https://doi.org/10.1080/13623699.2017.1382166.

Leveson, N., & Weiss, K. A. (2009). Software system safety. In *Safety design for space systems* (1st ed., pp. 475–505). Elsevier. https://doi.org/10.1016/B978-0-7506-8580-1.00015-4.

Maciejewski, M. (2017). To do more, better, faster and more cheaply: Using big data in public administration. *International Review of Administrative Sciences, 83*(1_suppl), 120–135. https://doi.org/10.1177/0020852316640058

Margetts, H., & Dorobantu, C. (2019). Rethink government with AI. *Nature, 568*(7751), 163–165. https://doi.org/10.1038/d41586-019-01099-5

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2022). A survey on bias and fairness in machine learning. *ACM Computing Surveys, 54*(6), 1–35. https://doi.org/10.1145/3457607

van Noordt, C., & Misuraca, G. (2022). Artificial intelligence for the public sector: Results of landscaping the use of AI in government across the European Union. *Government Information Quarterly, 39*(3), Article 101714. https://doi.org/10.1016/j.giq.2022.101714

van Noordt, C., & Tangi, L. (2023). The dynamics of AI capability and its influence on public value creation of AI within public administration. *Government Information Quarterly, 40*(4), Article 101860. https://doi.org/10.1016/j.giq.2023.101860

Overheid.nl. (2023). Het Algoritmeregister van de Nederlandse overheid. https://algoritmes.overheid.nl/.

Pérez López, C., Delgado Rodríguez, M., & de Lucas Santos, S. (2019). Tax fraud detection through neural networks: An application using a sample of personal income taxpayers. *Future Internet, 11*(4), 86. https://doi.org/10.3390/fi11040086

Rasmussen, J. (1997). Risk management in a dynamic society: a modelling problem. *Safety Science, 27*(2–3), 183–213. https://doi.org/10.1016/S0925-7535(97)00052-0

Rodriguez Rivas-Stellaard, S. D. (2023). *Boemerangbeleid: Over aanhoudende tragiek in passend onderwijs- en jeugdzorgbeleid*. Vrije Universiteit Amsterdam. https://doi.org/10.5463/thesis.141

Roland, H. E., & Moriarty, B. (1990). System safety engineering and management. *System Safety Engineering and Management*. https://doi.org/10.1002/9780470172438

Sætre, A. S., & Van de Ven, A. (2021). Generating theory by abduction. *Academy of Management Review, 46*(4), 684–701. https://doi.org/10.5465/amr.2019.0233

Selbst, A. D., Boyd, D., Friedler, S. A., Venkatasubramanian, S., & Vertesi, J. (2019). Fairness and abstraction in sociotechnical systems. *Proceedings of the Conference on Fairness, Accountability, and Transparency, 59–68*. https://doi.org/10.1145/3287560.3287598

Shein, E. H. (2004). *Organizational culture and leadership* (3rd ed.). Jossey-Bass.

Straub, V. J., Morgan, D., Bright, J., & Margetts, H. (2023). Artificial intelligence in government: Concepts, standards, and a unified framework. *Government Information Quarterly, 40*(4), Article 101881. https://doi.org/10.1016/j.giq.2023.101881

Sun, T. Q., & Medaglia, R. (2019). Mapping the challenges of artificial intelligence in the public sector: Evidence from public healthcare. *Government Information Quarterly, 36*(2), 368–383. https://doi.org/10.1016/j.giq.2018.09.008

The House Committee on Transportation and Infrastructure. (2020). Final Committee Report: The design, development & certification of the Boeing 737 max. https://transportation.house.gov/imo/media/doc/2020.09.15. FINAL 737 MAX Report for Public Release.pdf.

Tweede Kamer der Staten-Generaal. (2023). *Groningers boven gas - Rapport parlementaire enquêtecommissie aardgaswinning Groningen*.

Wieringa, M. (2020). What to account for when accounting for algorithms: A systematic literature review on algorithmic accountability. In *Conference on fairness, accountability, and transparency (FAT* '20)* (pp. 1–18). https://doi.org/10.1145/3351095.3372833

Wirtz, B. W., Weyerer, J. C., & Geyer, C. (2019). Artificial intelligence and the public sector—Applications and challenges. *International Journal of Public Administration, 42*(7), 596–615. https://doi.org/10.1080/01900692.2018.1498103

Wirtz, B. W., Weyerer, J. C., & Kehl, I. (2022). Governance of artificial intelligence: A risk and guideline-based integrative framework. *Government Information Quarterly, 39*(4), Article 101685. https://doi.org/10.1016/j.giq.2022.101685

Zhou, Z.-H. (2021). Machine learning. In *Machine learning*. Springer Singapore. https://doi.org/10.1007/978-981-15-1967-3.

Zuiderwijk, A., Chen, Y., & Salem, F. (2021). Implications of the use of artificial intelligence in public governance: A systematic literature review and a research agenda. *Government Information Quarterly*. https://doi.org/10.1016/j.giq.2021.101577. March, 101577.

Jeroen Delfos is a PhD candidate at the Faculty of Technology, Policy, and Management and is a data scientist at the Inspectorate of Justice and Security in the Netherlands. In his research, he uses empirical methods to study the impact of machine learning in the public sector. He aims to make the challenges of machine learning more explicit, making these insights valuable for scholars and policy makers alike.

Anneke Zuiderwijk is an associate professor in the Information and Communication Technology section of the Faculty of Technology, Policy, and Management at Delft University of Technology. Her research focuses on the reuse of governmental, research, and other data through infrastructures in open or more restricted forms. Anneke's research concerns data use, infrastructure functionalities that facilitate data use, (open) data ecosystems, socio-technical open data infrastructures and platforms, data publication, metadata, (open) data business models, and policy analysis.

Sander van Cranenburgh is an Associate Professor in the Transport and Logistics Group of the Technology, Policy, and Management faculty of Delft University of Technology. His research focuses on choice behavior analysis. His current research increasingly focuses on enriching the understanding of (travel) choice behavior through developing new data-driven modeling approaches. In this research, he specifically seeks the edge between traditional theory-driven approaches, such as Discrete Choice Models, and data-driven approaches, such as Artificial Neural Networks.

Caspar Chorus is the dean of the faculty of Industrial Design Engineering and professor of choice behavior modeling. His research aims to make choice models more realistic in terms of behavior without compromising on econometric elegance and usability. He developed the 'random regret minimization' model, which is widely used as an alternative to more traditional 'utility maximization' models.

Roel Dobbe is an assistant professor in the Information and Communication Technology section of the Faculty of Technology, Policy, and Management at Delft University of Technology. His research aims to create insight and actionable perspectives on how data-driven, learning-based and intelligent technologies can be integrated in safe and responsible ways, and how harm can be actively prevented. Roel has a PhD in Electrical Engineering and Computer Sciences from the University of California Berkeley (2018) and a MSc in Systems and Control from Delft University of Technology (2010). He is an active contributor to the establishment of governance practices for algorithmic and AI systems in public organizations, including in public administration, energy systems and healthcare.