Systemic Risks of Algorithms in Social Welfare: Combining Actor Analysis and System Safety Analysis

Wybe Segeren[1], Haiko van der Voort[1] and Roel Dobbe[1,*]

[1]Faculty of Technology, Policy and Management, Delft University of Technology, Jaffalaan 5, 2628BX, Delft, The Netherlands

Abstract

This paper considers the methodological challenge of mapping systemic risks and harms emerging from the use of algorithms and artificial intelligence in social welfare systems. Recent tragedies in social welfare put focus on the role of algorithmic systems in the execution of social security policies, motivating new governance and regulatory measures. While many efforts have tried to address risks at the level of the technology, individual process or organization, algorithmic risks in social welfare are inherently sociotechnical. Addressing these risks involves various actors, bringing in additional normative and political complexity. In this study, we apply two methods known for their ability to address parts of the complexity. Actor analysis is used to analyse the multi-actor aspect and associated normative dimensions, and a system safety analysis is used to map and analyze the sociotechnical nature and mechanisms of algorithmic risks. We motivate why and how these methods are combined and reflect on their synergy and challenges. The study is situated in the establishment of a Dutch algorithm watchdog, and focuses on the case of Dutch social security. As such, this study is a first of its kind to apply system safety to algorithms in the social welfare domain, and provides methodological contributions by using actor analysis to better scope and inform the multi-actor and cross-organizational nature of the safety analysis.

Keywords

social security, algorithmic systems, artificial intelligence, systemic risk, supervision, system safety, actor analysis

1. Introduction

The social domain has digitalized in recent years, seeking to benefit from the promise of efficiency, neutrality and ease for citizens [1, 2]. In the Netherlands, the execution of social security policies now uses algorithms and automated decision making at scale [3]. And while automation is often touted as a way of improving effectiveness and efficiency of human processes, science shows that reliably and safely depending on it in complex administrative contexts requires many new tasks and responsibilities [4, 5, 6]. The use of algorithmic systems, and more recently artificial intelligence (AI), has led to risks for society, which materialized in several cases in which harm was done to citizens [7, 5, 8]. The most notable example in the Netherlands is the childcare benefits scandal (Dutch: "toeslagenschandaal"), in which algorithms were used [9]. This led to increased attention for potential algorithmic harms in the social domain [10, 11]. In turn, this resulted in policy actions such as a public algorithm register, several assessment frameworks, and the founding of an algorithm watchdog [12, 13]. This watchdog is tasked with analyzing systemic risks of algorithms and AI, and coordinating with other supervisory authorities in the digital domain [12, 14]. Internationally, increased attention for algorithmic systems is visible in the EU AI Act, which recently went into force, bringing new measures for the use of AI in Europe.

Organizations using algorithmic systems and the organizations overseeing their use now find themselves in the middle of these national and international movements. They have a need to understand how the use of algorithmic systems can lead to harms, but also how these harms can be prevented. Merely having a technical understanding is not sufficient, as the systems and their effects are sociotechnical [15, 16, 17], and influenced heavily by political actors. Furthermore, while recent studies have showed the impacts of algorithmic tools in individual organizations [18, 19], many social welfare systems stretch

EGOV-CeDEM-ePart conference, August 31 - September 4, 2025, University for Continuing Education, Krems, Austria

*Corresponding author.

w.segeren@autoriteitpersoonsgegevens.nl (W. Segeren); h.g.vandervoort@tudelft.nl (H. v. d. Voort); r.i.j.dobbe@tudelft.nl (R. Dobbe)

0009-0007-3720-7709 (W. Segeren); 0000-0002-2795-9444 (H. v. d. Voort); 0000-0003-4633-7023 (R. Dobbe)

beyond the boundaries of single organizations and are multi-actor in nature. The factors involved in harms, and the associated intervention points, thus also stretch beyond boundaries of organizations [20, 21, 10]. These aspects lead to challenges for executive and supervisory agencies, and calls for methods that can be used to analyse and understand the social impacts and harms of algorithms used in complex, multi-actor sociotechnical systems.

A system safety approach looks at the safety of sociotechnical systems, and aims to analyze these as a whole. It considers safety as an emergent property, to be understood as the result of technical components, social behaviors and institutional aspects of a broader system [22]. This approach is therefore an apparent candidate for dealing with safety risks in complex social security systems [5]. However, system safety has historically dealt with clear understandings and consensus of what is safe and what is not, such as in aviation or nuclear safety [23, 24]. In many new domains, such consensus is not readily available and the political nature of a domain may make it hard to formulate clear system safety constraints [25]. What is safe and what is not can for instance be politically contested [26]. This is also the case for the use of algorithmic systems in the social welfare domain. System safety pioneer Nancy Leveson acknowledges the integration of these normative dimensions in system safety analysis as crucial but understudied [27]. Futhermore, even if consensus exists, complexity of the social domain makes it difficult to apply safety analysis straightforwardly, as the boundary of analysis spans multiple organizations and actors who are involved in designing and executing core functions.

On the other hand, actor analysis methods that allow understanding complex problems spanning multiple actors and organizations are available to help teasing out the complex normative and political dimensions of a problem [28, 29]. However, these typically lack the expressive abilities to display context-specificity and dynamics of how unsafe situations and harms emerge. While we are witnessing an uptick in the need to study the broader dynamics of AI safety across actors and organizations [30, 6], there are no studies that combine actor and safety analysis explicitly. For instance, Birkstedt et al. [31] point to a lack of attention in research for the sociopolitical and multi-actor context of governance of AI, and a lack of understanding of organization-level governance processes. The authors state that 'more research is needed to clarify organizational AI-governance processes and actor roles', with attention to both social and technical aspects, as well as their interactions.

In this paper, we study the combination and integration of actor analysis and system safety analysis in the context of algorithm supervision. The study was uniquely situated in the buildup of a new algorithm supervisory agency in The Netherlands, which focusses on systemic risks of algorithmic systems, including those in social welfare. As such, this study is further motivated by the practical question of how actor and system safety analysis can be combined to support emerging supervisory responsibilities for algorithms and AI in high-risk contexts.

The paper first provides more background on algorithm supervision in The Netherlands in Section 2, and motivates why a system-theoretic and multi-actor approach is suitable for conceptualizing systemic risks of algorithms in social welfare. Section 3 introduces the key methods and how these are applied on this case. Sections 4 and 5 then report insights from respectively the actor analysis and the system-theoretic process analysis. In Section 6 we share the results of our analysis and reflection on the combination of the methods. Section 7 provides conclusions and opportunities for future research.

2. Background and Theoretical Nature of Systemic Risks in Algorithmic Social Welfare

Section 2.1, covers relevant history of harms in social welfare systems in the Netherlands, and the advent of governance for algorithmic systems. Subsequently, Section 2.2 motivates and conceptualizes the theoretical nature of systemic algorithmic risks.

2.1. Development of a national algorithm supervisory landscape

The provision of social security policies in the Netherlands has digitalized, turning organizations executing social welfare policies into 'decision making factories' [32]. They make extensive use of algorithmic systems and are connected to each other through data exchanges [21]. Many Dutch social security policies are executed by municipalities together with two executive agencies: the Unemployment Insurance Agency (UWV) and the Social Insurance Bank (SVB). Their responsibilities are stipulated in the SUWI-law (Dutch: 'Wet Structuur Uitvoeringsorganisatie Werk en Inkomen'). The different involved parties automatically exchange information through a system called SUWInet, which also links to other organizations. UWV and SVB are not only responsible for executing social policies, but also for enforcing rules and preventing abuse. In both these roles, they make use of algorithms [32].

The executive agencies find themselves in a rapidly evolving context. Since the nineties, they were pushed by political actors to be more efficient, focus on fraud, and use new possibilities of data analysis and data exchanges [11]. Cases involving algorithmic harm resulted in attention for the safe use of algorithms in the public sector as well as increased focus on 'the human dimension' [11, 32, 33]. In social security, mistakes by algorithmic systems are impactful, and can directly impacts people's livelihoods. The executive agencies have taken action to work towards safe use of algorithms. At the same time, they have to relate themselves to increasingly complex social security policies, also adding complexity to the processes and algorithms used to execute them [3].

Scandals involving algorithms also resulted in action by the Dutch government. A strategy for value-driven digitalization was made, as part of which a public algorithm register for the public sector was launched, and an 'algorithm watchdog' was appointed. This task was appointed to Autoriteit Persoonsgegevens (AP), which is the Data Protection Authority also overseeing the EU General Data Protection Regulation (GDPR) in the Netherlands. Within AP, the new task was placed within the Department for the Coordination of Algorithmic Oversight (DCA). This authority was to act as a coordinating supervisor, working with sectoral supervisory bodies, as the use of algorithms stretches throughout different sectors, and risks are often sector- and context-specific [12, 34]. In addition, DCA was also tasked with analyzing and sharing systemic risks of Algorithms and AI, with specific attention for transparency, arbitrariness, and discrimination [12]. Recently, DCA has also worked on preparing the enforcement of the EU Artificial Intelligence (AI) Act together with other supervisory agencies, advising the government to formalize a coordinative approach for the AI Act, in which sectoral supervisors are involved, with DCA in a coordinating role together with the Dutch Authority for Digital Infrastructure (RDI) [35].

Despite efforts to promote safe use of algorithmic systems, in practice organizations still struggle to prevent recurrence of algorithmic harms. Furthermore, a multitude of developed frameworks for algorithm governance can be hard to filter through and use. And while the frameworks, legislation (e.g. GDPR, AI Act), and other efforts are steps towards safer development and use of algorithmic systems, these regulatory efforts are still in development and do not offer a golden bullet [31]. This is further complicated by recent geopolitical developments that are calling for deregulation of AI [36].

2.2. Need for a system-theoretic and multi-actor approach to mapping systemic risks

The aforementioned shows aspects that make a system-theoretic approach combined with actor analysis necessary. Algorithms are not standalone technical systems, but rather embedded in the context of social security. As such, they are part of sociotechnical systems that are structured by institutions, organizations and human professionals [31]. Systemic risks of algorithms, as evidenced in the childcare benefit scandal, are understood as *emergent system phenomena* that can only be understood by studying social components, technical components, and their interactions [27]. The background in 2.1 also shows that both algorithm governance and the social welfare system have their own history in the Netherlands, each intricately shaped by (ongoing) political dynamics and decisions.

Furthermore, the Dutch social welfare system is a multi-actor affair, with multiple executive organizations, various actors representing eligible citizens (e.g. client councils, ombudsmen), and different

authorities supervising the execution of policies. Supervision over algorithms in the Netherlands has also become a multi-actor problem, reflected in the coordinating role of DCA and its collaboration with other sectoral authorities. Actor Analysis provides a method to analyse the multi-actor aspect of problems at the actor level, and the level of the networks several actors shape [29]. It is used to provide an initial exploration of a multi-actor problem [28].

The multi-actor nature and political dynamics of the safety of algorithm use in social security makes it difficult to define clear and actionable definitions safety. Actors face a 'wicked problem' [37]. As defined by John Alford & Brian Head [38], these are complex, multifaceted issues that are resistant to resolution due to their evolving nature, conflicting interests, and interconnected causes. They often arise in public policy and governance, where no single solution satisfies all stakeholders. Alford & Head emphasize that wicked problems exhibit ambiguity, uncertainty, and contestation, making them difficult to define and address. Unlike "tame" problems, with clear solutions, wicked problems involve multiple stakeholders with diverse perspectives, each proposing different approaches based on their values and interests. Another key characteristic is dynamism - the problem itself can change over time, so solutions must be adaptive and iterative. Furthermore, solutions to wicked problems often have unintended consequences, sometimes creating new problems rather than resolving the original one.

The wicked nature of this issue complicates application of STPA, and asks for analysis methods that allow for differing conceptions of safety. Actor Analysis looks into differing values and perceptions of actors [28, 29]. As such, the central problem of mapping systemic risk of algorithms in the Dutch social welfare system can be conceptualized by combining theory from policy analysis for multi-actor systems [28] and concepts from system safety for complex systems subject to software-based automation and AI [5]. In the following section, we outline the core associated methods and how we combine these.

3. Methodology

This section introduces key methods and how these were applied. For further details about the methods and methodology, we direct the reader to the associated thesis [39]. This work takes a case-based approach, applying Actor Analysis and System-Theoretical Process Analysis (STPA) to the case described in Section 2.1. This multi-method approach aims to build on strengths of both methods, but also address their weaknesses, in two ways. First, to provide insight into the mechanisms in the use of algorithmic systems in execution of social security policies that can lead to systemic harms, also clarifying how these can be addressed across networks of actors. And second, to provide insight into the efficacy and limitations of the two methods for analyzing similar sociotechnical systems in novel domains.

The first analysis is an Actor Analysis (AA), described by Enserink et al. [28], which is a collection of analyses that together shape a view of the actor networks surrounding a problem situation, the perceptions the actors have, the values they uphold, and their resources, thereby providing a detailed insight into the sociopolitical context of the case. Section 4 provides more detail about the steps of AA.

Data is gathered using document analysis and semi-structured interviews. These form a typical foundation under AA [29]. The extensive document analysis includes academic literature and diverse gray literature, including government documents, organizational reports, organizational plans, and other policy documents. Semi-structured interviews with experts in the field were held, and analyzed by using coding, providing rich insight into granular aspects including issues, considerations, and operational processes in practice. The interviews were conducted alongside execution of AA, which helped validate the scope and results of the Actor Analysis and document analysis. Further information on data sources and interviews can be found in the associated thesis [39].

AA and the sourced data were used as a basis of knowledge to build the subsequent STPA. This method was developed by Leveson [27], and analyses safety as an emergent property of sociotechnical systems. STPA is based on a model of the core operational process in which algorithms and automation are integrated, as well as of the adjacent processes, including management, maintenance, system design, policy and enforcement. Together, these processes and their interactions comprise a *safety control structure* which integrates all the various duties and interventions across processes and actors that are

relevant to ensuring the safety of the operational process. These structures allow to build understanding of hazards that can arise in these processes as result of inadequacies. The specific steps followed to analyze algorithmic harm through STPA are outlined in Section 5.

As mentioned, the methods were applied to the case of social security policies in the Netherlands. The two agencies that execute these policies use algorithmic systems for awarding benefits and enforcing rules. At the same time, supervision over the use of algorithmic systems is under development, with a supervisory authority for algorithms and AI coordinating supervisory actions and analyzing systemic risks. This case is further illustrated in the Actor Analysis.

4. Actor Analysis

Actor Analysis, as described by Enserink et al. [28], consists of six steps that can be divided into two parts. The first part analyses the actor network surrounding a problem situation. The second part then dives further into the actor-level. This section is structured following these two parts.

4.1. Problem situation, actors, and formal relations

The central problem situation in this analysis is the use of algorithmic systems in executing social security policies can lead to harms for citizens. Figure 1 shows the formal chart with the involved actors and the formal relations between them. The multitude of included actors shows the complex nature of the case. Actors are related through formal relations, but also through displayed cooperations. Several actors are related through SUWInet, a network for sharing data between different executive agencies, intricately linking their processes and data.

The involvement of algorithmic systems adds to the complexity of the actor field. Not only are actors in the field of social security involved, so are multiple supervisory agencies and ministries that focus on algorithms and AI. Algorithmic systems thus shape a complex policy problem: they can never be considered in their own right, but need to be viewed in their context of application, involving a plethora of actors with different roles and viewpoints. Actors focusing on algorithms need to do this for a multitude of application domains, of which social security is only one example.

Notable in 1 is the lack of formal powers for several actors. Interest groups have little formal power, and rely on informal ways of influencing the problem situation. The newly appointed coordinating supervisor also held no formal power at the time of analysis. It therefore has to rely on informal relations, and formal powers of other supervisory agencies through its coordinating role.

4.2. Looking further at actors: Actor characteristics and potential cooperation

Next, the actors related to the problem situation are analyzed further. A first step looks at the characteristics of the individual actors: their interests, objectives, perceptions and resources. It found important differences of viewpoints between actors, mostly relating to differences in formal roles of actors, such as differences in focus between ministries. Notably, organisations struggle with their own ambivalences towards the impact of algorithms and the tradeoffs between risks and efficiencies of their use.

Such ambivalences were present in the executive agencies, complicating their ability to handle the situation. This relates to what Kuziemski & Misuraca [40] describe as a 'tragic double bind'. The agencies need algorithms to execute policy efficiently, but also need to ensure this does not impose harm on citizens. They need to 'govern by algorithms, but also govern algorithms' [40]. The analysis found three additional 'double binds'. First, organizations need to be careful and keep the 'human dimension' into account, but are also pushed to be efficient, use standardization and automate processes. Second, the organizations are responsible for serving citizens, but also for enforcing rules and taking action against fraud. And third, the agencies aim to find solutions that fit the needs of individual citizens, but on the other hand need to guarantee equality across citizens.

These double binds reflect different viewpoints of political actors and society. The two sides are sometimes hard to commensurate, asking for tradeoffs to be made. The result can shift over time,

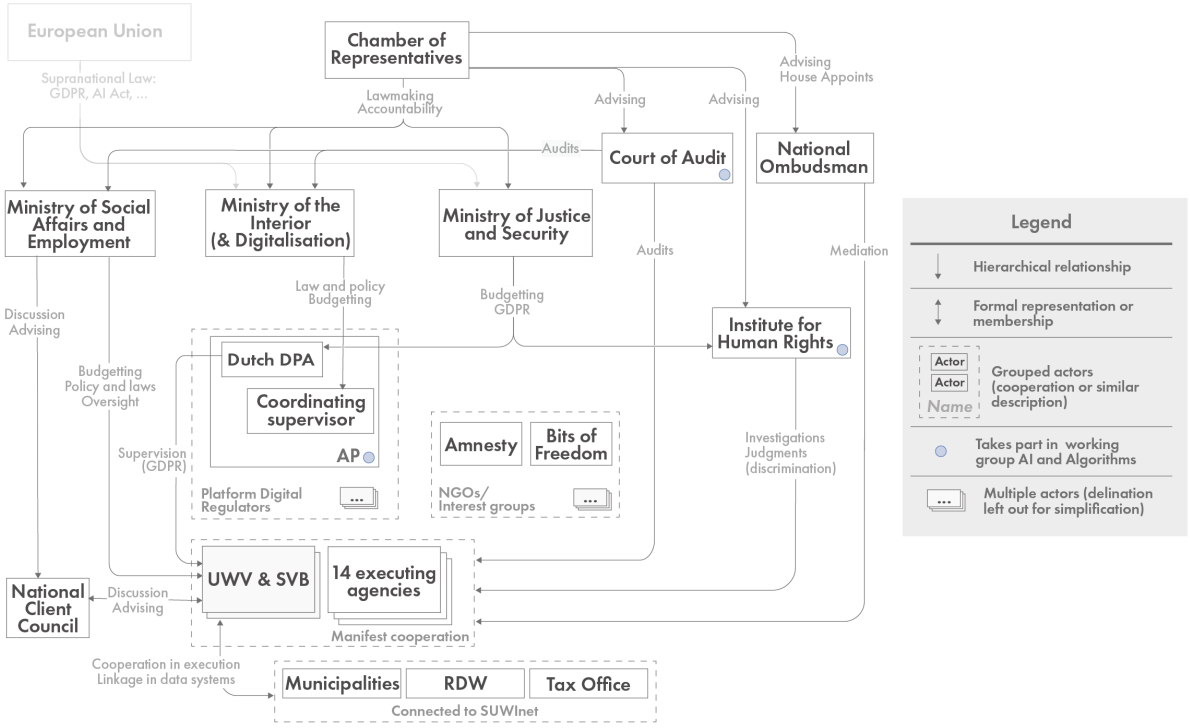

Figure 1: Formal Chart

impacted by the sociopolitical climate. In the Netherlands, a focus on efficiency and action against fraud in the early 2010s was instrumental in the onset of the childcare benefit scandal [11, 10], which in turn led to a shift of focus towards carefulness and serving citizens by the early 2020s. Recently, however, we see a resurfacing of focus on efficiency, framed around AI innovations and the need to address issues of labor productivity and shortage of personnel.

Analysis of the resources and dedication of actors is the next step and aims to identify the potential for cooperation on preventing and addressing algorithmic harm. This analysis found resource gaps: some dedicated actors currently hold little power to influence systemic risks of algorithms in social security. As mentioned, an important example is the coordinating supervisor, as displayed in the Formal chart (see Figure 1). This formal chart also shows that important potentials for cooperation are already fulfilled. Supervisory agencies work together around the topic of digitalisation, in the Platform Digital Regulators. Different executive agencies work together in the Manifest group. Such cooperation can form a basis for comprehensively addressing systemic risk of algorithms.

5. System-Theoretical Process Analysis (STPA)

This analysis builds on the outcomes of the Actor Analysis. It uses the same data sources: document analysis and semi-structured interviews with experts in the field. STPA uses the concepts *safety, losses, and hazards*. *Safety* is considered to be freedom of *losses*, which are those consequences that are unacceptable to stakeholders. They result from *hazards*, which are the system states that in worst-case conditions lead to accidents and losses. Safety is ensured by constraints on system components and their interactions, which together should prevent hazards from occurring, thereby preventing losses and keeping the system safe. The section is structured along the four steps of STPA [22, 24].

5.1. Purpose of the analysis

The starting point of STPA is defining the purpose of analysis, i.e. the system in focus, losses, hazards, and core constraints to be satisfied. The Actor Analysis provides a starting point, describing the system

in focus from an inter-organizational viewpoint. It also broadly described the safety issue: the use of algorithmic systems to execute social security policies can lead to harms for citizens and society. This issue can be described more specifically by unraveling important losses for citizens and society. The losses, listed in Table 1, were delineated using information from the interviews and Actor Analysis.

The hazards that can lead to these losses can be found in Table 2. These are general hazards, fitting to the wider viewpoint taken in this analysis. Delineating specific constraints was left out of scope, as this research focused on understanding the losses and hazards. Providing specific constraints furthermore proved difficult, both due to the type of system analyzed and the broad scope that was applied.

Table 1
Losses used in STPA

Identified losses
(1) Citizen unjustly uses benefits (maliciously/accidentally)
(2) Citizen does not use benefits they are entitled to
(3) Benefits are incorrectly appreciated
(4) Citizen is subjected to unjust or disproportionate investigations and measures
(5) Citizen is not efficiently helped
(6) Citizen is unable to exercise their rights

Table 2
Hazards and related losses

Hazards	Related losses
Flawed logic is used in algorithmic systems	3,4,5
Flawed information is used in algorithmic systems	1,3,4
Citizen has flawed understanding of role of algorithms in benefits system	1,2,6

5.2. Safety Control Structures

Safety control structures display controlled processes and controllers, with feedback and control actions between them: controllers control processes through control actions and receive feedback from the processes, allowing adjustment of control actions. Controllers can be human or machine-based or a combination of both. Typically, control diagrams display processes and controllers within one or a few organizations. In this research, the Actor Analysis allowed to construct a control structure that showed inter-organizational components. Scoping down, two control diagrams showing a single organization each were made for UWV and SVB. These show algorithmic systems used in processes and the organizational structure they are embedded within. Building control diagrams on multiple levels allows to analyze at the process level, organizational level, and inter-organizational level, potentially allowing to delineate how factors from the inter-organizational level (e.g. political factors) can lead to losses on the process level. The complexity of the safety control structures is too high to portray in this article. Instead, six defining characteristics are shared below. The interested reader may find further details in the associated thesis [39].

First, although the processes and algorithmic systems are diverse, there are archetypes in the ways algorithmic systems are used in processes, as well as the way these systems and the processes are embedded in organizations. Second, the information exchange system SUWInet is an important common factor for several processes, adding complexity to processes and directly connecting different organizations' data infrastructures. In addition, the organizations also still have several internal databases. Third, processes show a typical hierarchical structure: algorithmic systems are used by executing employees, under control of a manager, director, and the board. The board interacts with the Ministry of Social Affairs, and so on. Fourth, several divisions in UWV and SVB are relevant. Maintenance, supervision, development and validation are sometimes done by a separate division from the one using the algorithms, although cooperation happens. This complexity is also present in

the wider actor landscape, where actors focusing on algorithms come together with those focusing on social policies. Fifth, the divisions that work on algorithms interact with external parties, namely external auditors, advisory parties, but also an ethical committee and citizen panels or other citizen representation. Experts are also involved more closely within specific processes. Sixth, UWV and SVB organize fora to discuss issues related to algorithmic systems. One agency has a dedicated 'coalition' wherein management discusses the systems and associated policies. The other has workshops where employees from different divisions can discuss diverse issues they see in their work. These can be seen as different approaches to organizational governance of algorithms. One more formal and top down, the other flexible and bottom-up.

5.3. Identifying Unsafe Control Actions

Control diagrams allow the identification of *inadequate control* that can lead to hazards and losses, both in the operational process and the broader organizational system. Unsafe control actions (UCAs) can be social and technical, and can thus also include relevant political factors. Since an inter-organizational control diagram was made, factors outside the organizations can also be considered. Furthermore, it allows for analysis of the role of supervision of execution of policies and the use of algorithmic systems. This step works towards understanding the mechanisms of harms and how these emerge across a sociotechnical system. Our analysis identified three categories of UCAs.

UCAs in the processes wherein algorithms are used. These relate to decisions made by or using algorithms, the actions taken as a result of them, and the information that is given to citizens. Examples are calculations of benefits to receive, investigations, and measures.

UCAs within the executive organizations. These relate to processes beyond the immediate operational use of algorithms. These include control actions in development, maintenance, and the monitoring of models, their outputs and their use. This also includes consultation of external parties and citizen representation, as well as internal policy and instructions from management, both to executing employees and those making and monitoring the algorithms.

UCAs outside the executive organization. In the wider ecosystem, UCAs include the automated data transfers between government agencies, as well as law and policy creation by political actors, which influences supervision, social security, and internal policy and may determine unsafe outcomes.

5.4. Loss Scenarios

The UCAs are build out to loss scenarios, which combine both direct and indirect factors that contribute to the occurrence of unsafe control actions, hazards and finally losses. These scenarios provide insight into the sociotechnical mechanisms of algorithmic harm, and can inform various places to intervene for harm prevention, by addressing the various factors. Our analysis prioritized three main loss scenarios.

5.4.1. Scenario 1: Flawed logic used in algorithmic systems

Flawed logic used in algorithmic systems, such as selection criteria in enforcement algorithms and categorization of disabled people's ability to work, can cause harms, such as erroneous calculations of benefits and unjustified investigations and measures. The logic used is influenced by the development and monitoring of algorithms and related processes. Involvement of citizens can be important to understand the reality of policies. Narrow monitoring, such as strict use of KPIs, decreases the ability to flag hazards. Monitoring should therefore be open to a broad array of signals. Involving citizens can help adjust control actions in this respect [6, 41]. The scope of monitoring is influenced by a focus internally and externally on efficiency. Discretionary space for executive employees might help to prevent accidents. However, meaningful human intervention requires knowledge, ability, and freedom to act. This asks for adequate internal policy, processes, and training, all of which are influenced in turn by management instructions, ministries, and their laws and policies.

5.4.2. Scenario 2: Flawed information used in calculations using algorithms

Use of flawed information can cause similar hazards and losses as the use of flawed logic. Important factors include the complexity of social security policies and automated data exchanges through SUWInet. This exchange platform connects databases from multiple agencies, easing the execution of their tasks. However, with different organizations and different tasks, there are risks for differences in definitions, and for erroneous and untimely data. Changes in data can cause effects elsewhere in the system, sometimes automatically [42, 11]. Indirectly, coupling of data sources is pushed for by political actors. The complexity of couplings and effects is related to the complexity of social security policies. Solution directions include pushing for communication between actors that exchange data, common definitions, but also mapping of inter-dependencies between systems and policies.

5.4.3. Scenario 3: Citizens have flawed understanding of role of algorithms

If citizens do not know if or how algorithms were involved, this impacts their ability to attain their rights by contesting decisions or adressing sources of mistakes [32]. Lacking transparency and high complexity of algorithms and related policies make it hard for citizens to comprehend the social security system and the role of algorithms therein. In the supervisory landscape, this knowledge is currently not present to a sufficient extent either. Supervisors have limited means, formal powers, and there is limited central knowledge creation. Furthermore, algorithm registration is not obligated by law [9]. Political actors can steer the funds, focus, and legal abilities of supervisory actors, and decrease complexity of social policies and their execution. Structural knowledge of supervisors can help prevent losses. Transparency towards citizens should allow citizens to take meaningful actions [16].

6. Discussion

This work showed the complexity of the use of algorithms in the social welfare domain, and supervision thereof. It requires consideration of social and technical aspects, including embedding in operational processes and actor networks. Policies and actors focused on social welfare systems intersect with those focusing on algorithms and artificial intelligence. System-Theoretical Process Analysis (STPA) and Actor Analysis (AA) have the ability to map these complex system aspects, and together provide a more complete system mapping than either analysis on its own. As such, the combined methods have the potential to provide insights for supervisory actors and organizations using algorithms, both of whom have to relate to the complex system they are situated within, and the wicked problems they face. The methods can help bridge the gap between different actors and professions, in order to come shared understandings of the potential for algorithmic harm and work towards preventing it.

An important affordance of combining the two methods is that AA serves as a starting point on which the more granular STPA can be built. AA provides an initial overview of the system in focus and important considerations that stem from the actors in this system. The hierarchical approach of both methods aligns the methods. Furthermore, outcomes of AA can contextualise findings of STPA. For instance, harms might be explained by lacking means, viewpoints, or conflicts in the actor field. AA also guides the solution directions proposed to the causal scenarios under STPA. For instance, improvements might neccessitate knowledge about whether the actors can and want to execute them, together or alone. This means realistic options for improvement can be suggested and mapped. AA also widens the viewpoint of STPA, which typically shows one or few organizations. AA is inherently more inter-organizational, as it take a multi-actor viewpoint. This broader scope adds to the explanatory value of STPA, as it is able to consider a broader range of factor, including political and social factors.

On the other hand, a wider scope provides a challenge for STPA. It becomes harder to specify general hazards and losses for the overall system. Another challenge lies in scoping the analysis in this broader viewpoint. Mapping an inter-organizational system risks saturating the analysis with processes, controllers, control actions, and feedback. STPA includes more components than AA (i.e. more than just actors), and in a wider scope, even more components can be considered. Scoping efforts

thus become more important, which risks bringing subjectivity into the analysis. However, AA and interviews allow the researcher to consider what is important from the viewpoint of the actors, guiding the scoping efforts. Finally, while STPA lends itself for analyzing values other than safety [22], there are few studies available that do so. As such, Actor Analysis may move STPA beyond 'safety advocacy', as it respects multiple values as part of the analysis and explains where they come from.

7. Conclusion

The social welfare case discussed in this work illustrates how algorithmic harm is a multi-actor, socio-technical affair. This points to the complexity of supervising, mapping, and designing for safety in these systems. Actor Analysis (AA) showed the complexities of the multi-actor nature of the case and sketched algorithmic harm as a wicked problem. It highlighted limitations in formal means and ambivalences in goals of actors. AA was also a starting point for System-Theoretical Process Analysis (STPA), which showed a new way of looking at algorithms in their sociotechnical context. Through analysis of control diagrams and creation of loss scenarios, the mechanisms of algorithmic harms were mapped, and possible interventions identified. The multi-method approach shows potential in analysing complex sociotechnical systems in novel domains. AA provides a basis and additional explanatory value for STPA. Consideration of a broader actor field and the normative aspects of the sociopolitical context, adds to the ability of STPA to analyse algorithmic harm. However, it proved difficult to produce generic hazards and losses for a broad scope, and AA proved to more easily encorporate values than STPA.

It is important to work on providing constraints for novel domains, such as described in this work, where conceptions of safety might not be clear-cut. Building on this work, research can also seek to further incorporate values in system safety analyses. Different conceptions of what safety means could hold explanatory value for the dynamics of hazards and losses. Work on the safe use of algorithms finds itself in a rapidly evolving context. Future work should take into account continuous regulatory efforts, including consequences of the EU AI Act and related supervisory structures. It may also help to counter calls for early deregulation of AI. Both developments bring changes to the actor field and dynamics therein. The combined analysis methods presented in this research provide abilities to analyze such dynamics, and the resulting potential for algorithmic harm in the future, as well as how to prevent it and determine interventions and distribution of responsibilities.

Acknowledgments

Gratitude goes out to Autoriteit Persoonsgegevens (AP, the Dutch Data Protection Authority) for providing opportunities for empirical research, in particular Gerald Hopster and Stefan Kulk for their contribution to the thesis supervision.

Wybe Segeren was employed as a research intern and received research support from AP. He is currently employed by the Department for the Coordination of Algorithmic Oversight (DCA) in the Netherlands. The research project was completed prior to this employment. This work does not reflect views and opinions of the organization.

Declaration on Generative AI

The authors have not employed any Generative AI tools.

References

[1] C. Prins, H. Sheikh, E. Schrijvers, E. de Jong, M. Steijns, M. Bovens, Opgave AI: De nieuwe systeemtechnologie, , Wetenschappelijke Raad voor het Regeringsbeleid (WRR), Den Haag, 2021.

[2] M. Raub, Bots, Bias and Big Data: Artificial Intelligence, Algorithmic Bias and Disparate Impact Liability in Hiring Practices, Arkansas Law Review 7 (2018) 529–570.
[3] Tijdelijke Commissie Uitvoeringsorganisaties, Klem tussen balie en beleid, , Tweede Kamer der Staten-Generaal, Den Haag, 2021. URL: https://www.tweedekamer.nl/kamerleden-en-commissies/commissies/tijdelijke-commissie-uitvoeringsorganisaties/eindrapport.
[4] I. Lindgren, Ironies of automation and their implications for public service automation, Government Information Quarterly 41 (2024) 101974. doi:10.1016/j.giq.2024.101974.
[5] R. Dobbe, System Safety and Artificial Intelligence (2022). doi:10.48550/arXiv.2202.09292.
[6] J. Delfos, A. M. G. Zuiderwijk, S. van Cranenburgh, C. G. Chorus, R. I. J. Dobbe, Integral system safety for machine learning in the public sector: An empirical account, Government Information Quarterly 41 (2024) 101963. doi:10.1016/j.giq.2024.101963.
[7] Autoriteit Persoonsgegevens, AP Inzet Artificial Intelligence Act, , 2022. URL: https://www.autoriteitpersoonsgegevens.nl/documenten/ap-inzet-artificial-intelligence-act-ai-act.
[8] M. Wieringa, "Hey SyRI, tell me about algorithmic accountability": Lessons from a landmark case, Data & Policy 5 (2023) e2. doi:10.1017/dap.2022.39.
[9] E. Nieuwenhuizen, Algorithm registers: A box-ticking exercise or meaningful tool for transparency?, Information Polity 29 (2025) 415–433. doi:10.1177/15701255241297107.
[10] R. Peeters, A. C. Widlak, Administrative exclusion in the infrastructure-level bureaucracy: The case of the Dutch daycare benefit scandal, Public Administration Review 83 (2023) 863–877. doi:10.1111/puar.13615.
[11] W. van Atteveldt, W. Roozendaal, N. Ruigrok, K. de Vries, M. van der Velden, M. Busuioc, P. Guldemond, Tussen Ambitie en Uitvoering. Een contextanalyse van de dynamiek tussen media, politiek en beleid bij de totstandkoming en uitvoering van dertig jaar sociale zekerheid, , Vrije Universiteit Amsterdam, 2024. URL: https://hdl.handle.net/1871.1/a6e17f7e-7fec-49c1-b9d0-00bfaface3e8.
[12] A. C. van Huffelen, Inrichtingsnota Algoritmetoezichthouder, kamerbrief kst-26643-953, Tweede Kamer der Staten Generaal, Den Haag, 2022.
[13] Handreiking Algoritmeregister: aan de slag met het Algoritmeregister, , Ministerie van Binnenlandse Zaken en Koninkrijksrelaties, 2023. URL: https://www.digitaleoverheid.nl/document/handreiking-algoritmeregister/.
[14] A. C. van Huffelen, Opschaling Algoritmetoezichthouder, kamerbrief 2023-0000738933, Ministerie van Binnenlandse Zaken en Koninkrijksrelaties, Den Haag, 2023.
[15] A. Balayn, S. Gürses, Beyond Debiasing: Regulating AI and its inequalities, , EDRi, 2021. URL: https://edri.org/wp-content/uploads/2021/09/EDRi_Beyond-Debiasing-Report_Online.pdf.
[16] N. Diakopoulos, Transparency, in: The Oxford Handbook of Ethics of AI, Oxford University Press, 2020, pp. 197–213. doi:10.1093/oxfordhb/9780190067397.013.11.
[17] J. Hamer, A. Lemmens, L. Kool, Algoritmes Afwegen: Verkenning naar maatregelen ter bescherming van mensenrechten bij profilering in de uitvoering, , Rathenau, Den Haag, 2022.
[18] D. Toll, I. Lindgren, U. Melin, Stakeholder Views of Process Automation as an Enabler of Prioritized Value Ideals in a Swedish Municipality, JeDEM - eJournal of eDemocracy and Open Government 14 (2022) 32–56. doi:10.29379/jedem.v14i2.726, number: 2.
[19] S. Nouws, M. Janssen, R. Dobbe, Dismantling Digital Cages: Examining Design Practices for Public Algorithmic Systems, in: Electronic Government, Springer, Cham, 2022, pp. 307–322. doi:10.1007/978-3-031-15086-9_20, iSSN: 1611-3349.
[20] A. Widlak, R. Peeters, Administrative errors and the burden of correction and consequence: how information technology exacerbates the consequences of bureaucratic mistakes for citizens, International Journal of Electronic Governance 12 (2020) 40–56. doi:10.1504/IJEG.2020.106998.
[21] R. Peeters, A. Widlak, The digital cage: Administrative exclusion through information architecture – The case of the Dutch civil registry's master data management system, Government Information Quarterly 35 (2018) 175–183. doi:10.1016/j.giq.2018.02.003.
[22] N. Leveson, J. Thomas, STPA Handbook, 2018. URL: http://www.flighttestsafety.org/images/STPA_Handbook.pdf.
[23] N. Leveson, Safety III: A Systems Approach to Safety and Resilience (2020). URL: https://psas.

scripts.mit.edu/home/nancys-white-papers/.
[24] S. Rismani, R. Dobbe, A. Moon, From silos to systems: Process-oriented hazard analysis for ai systems, 2024. URL: https://arxiv.org/abs/2410.22526. arXiv:2410.22526.
[25] H. G. van der Voort, A. J. Klievink, M. Arnaboldi, A. J. Meijer, Rationality and politics of algorithms. Will the promise of big data survive the dynamics of public decision making?, Government Information Quarterly 36 (2019) 27–38. doi:10.1016/j.giq.2018.10.011.
[26] T. Swierstra, P. Vermaas, The Entanglement of Technology and Morality, in: T. Swierstra, P. Lemmens, T. Sharon, P. Vermaas (Eds.), The Technical Condition: The Entanglement of Technology, Culture, and Society, Uitgeverij Boom, Amsterdam, 2022, pp. 239–268.
[27] N. Leveson, Engineering a Safer World: Systems Thinking Applied to Safety, The MIT Press, Cambridge, Massachusetts London, England, 2012. doi:10.7551/mitpress/8179.001.0001.
[28] B. Enserink, P. Bots, E. van Daalen, L. Hermans, R. Kortmann, J. Koppenjan, J. Kwakkel, T. Ruijgh-van der Ploeg, J. Slinger, W. Thissen, Policy Analysis of Multi-Actor Systems, volume 2, TU Delft Open, 2022. doi:10.5074/T.2022.004.
[29] L. Hermans, W. Thissen, Actor analysis methods and their use for public policy analysts, European Journal of Operational Research 196 (2009) 808–818. doi:10.1016/j.ejor.2008.03.040.
[30] Y. Bengio, S. Mindermann, D. Privitera, T. Besiroglu, R. Bommasani, S. Casper, Y. Choi, P. Fox, B. Garfinkel, D. Goldfarb, H. Heidari, A. Ho, S. Kapoor, L. Khalatbari, S. Longpre, S. Manning, V. Mavroudis, M. Mazeika, D. Michael, (...), Y. Zeng, International AI Safety Report, 2025. doi:10.48550/arXiv.2501.17805.
[31] T. Birksteds, M. Minkkinen, A. Tandon, M. Mäntymäki, Ai governance: themes, knowledge gaps and future agendas, Internet Research 33 (2023) 133–167. doi:10.1108/INTR-01-2022-0042.
[32] D. Houtzager, S. Verbeek, A. Terlouw, Gelijk recht doen Deelrapport Sociale Zekerheid (2022). URL: https://www.eerstekamer.nl/overig/20220614/gelijk_recht_doen_deelrapport_3/meta.
[33] Stand van de uitvoering, Technical Report, Ministry of Social Affairs, The Netherlands, 2023. URL: https://www.rijksoverheid.nl/documenten/kamerstukken/2024/01/22/kamerbrief-stand-van-de-uitvoering-december-2023.
[34] S. Nas, S. Ouburg, Inrichting Algoritmetoezicht: Scenario's korte en lange termijn, Advisory Report, Privacy Company, 2022. URL: https://www.rijksoverheid.nl/documenten/rapporten/2022/12/05/inrichting-algoritmetoezicht.
[35] M. Verdier, A. van Dijk, Eindadvies inrichting AI-toezicht AP & RDI, 2024. URL: https://www.autoriteitpersoonsgegevens.nl/documenten/eindadvies-inrichting-ai-toezicht-ap-rdi.
[36] R. Dobbe, AI Safety is Stuck in Technical Terms – A System Safety Response to the International AI Safety Report, 2025. doi:10.48550/arXiv.2503.04743.
[37] A. Zuiderwijk, Y.-C. Chen, F. Salem, Implications of the use of artificial intelligence in public governance: A systematic literature review and a research agenda, Government Information Quarterly 38 (2021). doi:10.1016/j.giq.2021.101577.
[38] J. A. . B. Head, Wicked ans less wicked problems: a typology and a contingency framework, Policy and Society 3 (2017) 397–413.
[39] W. Segeren, Governing Algorithmic Systems in the Social Domain, Published Master's Thesis, Delft University of Technology, 2024. URL: https://resolver.tudelft.nl/uuid:d8c6a404-7a9d-4814-8e15-b8af760fe816.
[40] M. Kuziemski, G. Misuraca, AI governance in the public sector: Three tales from the frontiers of automated decision-making in democratic settings, Telecommunications Policy 44 (2020) 101976. doi:10.1016/j.telpol.2020.101976.
[41] S. Grimmelikhuijsen, A. Meijer, Legitimacy of Algorithmic Decision-Making: Six Threats and the Need for a Calibrated Institutional Response, Perspectives on Public Management and Governance 5 (2022) 232–242. doi:10.1093/ppmgov/gvac008.
[42] A. Widlak, Overzicht & Ondergrens: digitale overheid in kort bestek, Technical Report, Stichting Kafkabrigade, 2022. URL: https://staatvandeuitvoering.nl/onderzoek/overzicht-ondergrens-digitale-overheid-in-kort-bestek/.