W.W.C. (Wybe) Segeren

Governing Algorithmic Systems in the Social Domain

Providing a sociotechnical description of algorithmic systems in the SUWI-domain using Systems Safety and Actor Analyses

Master Thesis
MSc Complex Systems Engineering and Management
Faculty of Technology, Policy, and Management

June 17, 2024
Wybe Willebrord Christiaan Segeren
Student Number: 4681231

This work was made in cooperation with Autoriteit Persoonsgegevens, Department for the Coordination of Algorithmic Oversight

AUTORITEIT
PERSOONSGEGEVENS

TU Delft

Graduation Committee

Dr. Ir. R.I.J. (Roel) Dobbe (TU Delft)
Dr. H.G. (Haiko) van der Voort (TU Delft) (Chairperson)
G. (Gerald) Hopster (Autoriteit Persoonsgegevens)
Mr. Dr. S. (Stefan) Kulk (Autoriteit Persoonsgegevens)

Executive Summary

Attention for algorithmic systems (AS), including Artificial Intelligence, has increased in recent years. Public organisations have digitalised, using standardisation and AS to automate tasks and provide public services more easily and efficiently (Hamer et al., 2022; Tijdelijke Commissie Uitvoeringsorganisaties, 2021). The flipside of these developments is the potential for harms stemming from AS. Automated systems can increase burdens for citizens, cause exclusion, and lead to discrimination (Balayn & Gürses, 2021; Peeters & Widlak, 2023; Raub, 2018). Furthermore, they fundamentally change the workings of government, and pose threats for legitimacy, accountability and transparency (Grimmelikhuijsen & Meijer, 2022; Widlak, 2022; Wieringa, 2023; Zouridis et al., 2019). The potential for harms has already materialised in the Netherlands, with cases surrounding anti-fraud system SyRI, the childcare benefits scandal, and the recent DUO-case as examples (Prins, 2021; Vié, 2023; Wieringa, 2023). These cases involve executing agencies, governmental organisations that execute policy. They increasingly use AS to execute policies and enforce related rules. Following the potential and materialisation of risks stemming from AS, government and supervisory authorities have taken action. It led to the instalment of a national algorithm coordinator, the Department for the Coordination of Algorithmic Oversight (DCA), as well as a national algorithm register for government, and a plethora of guidelines and frameworks.

Complex socio-technical systems are compositions of intertwined social and technical components, mutually influencing each other but never fully determining each other (Bauer & Herder, 2009). Such intertwinement is also present in algorithmic systems, as they too are complex socio-technical systems (Wieringa, 2020). These complex systems lead to non-linear and non-predictable outcomes (Sargut & McGrath, 2011). While AS are socio-technical, regulatory responses can have a technocentric focus (Balayn & Gürses, 2021). Socio-technical analyses providing overview of use and governance of AS, including recent policy actions, can help understand how safe AS-use is, and what can be done to improve governance. Currently, such an overview is missing, while socio-technical analyses, systems safety in particular, can provide valuable lessons for the field (Dobbe, 2022). This study aimed to analyse the governance of AS at executing agencies executing social security policies surrounding work and income (SUWI). Under this law, two executing agencies, UWV and SVB, execute policy using AS. The aim of this study was to answer the following research question:

How can a systems safety perspective combined with an actor perspective be used to provide a sociotechnical description of algorithmic systems, their governance and possible hazards at agencies executing the SUWI-law?

An Actor Analysis (AA) was done, as described by Enserink et al. (2022). It looks into actors' viewpoints on safety of AS in SUWI, formal roles and relations, possible solutions actors propose, power, interest and possible actor coalitions. The analysis found that DCA currently has no formal abilities, relying on informal methods only. Current actor cooperations fulfil the potential for cooperations in governance of AS. The analysis found ambivalences within and between actors, described as 'double binds', following Kuziemski and Misuraca (2020). Executing agencies need to govern with AS while also governing AS. Furthermore, there are ambivalences between carefulness versus efficiency, serving citizens versus strong action against fraud, and finding fitting solutions versus guaranteeing equality.

Next, a System-Theoretical Process Analysis (STPA) was executed, following Leveson and Thomas (2018). STPA moves on from ensuring safety through improving reliability of system components, stating safety requires designing the entire socio-technical system so that it remains safe under varying conditions. In describing safety, it uses descriptions of hazards, which are system states that lead to losses. Losses are situations with undesired effects for stakeholders. STPA uses a hierarchical Safety Control Structure (SCS), displaying processes and those controlling them, as well as control actions and feedback between them. Hazards stem from inadequate control actions and feedback, and causal scenarios are used to understand how hazards can arise in the SCS. The analysis focussed on three hazard categories. First, the use of flawed logic in AS was related to limited insight into complex realities of citizens being translated into AS and their monitoring. Second, the use of flawed information in AS was coupled to frequent data exchanges between organisations, where differences in use cases and data quality can cause issues for citizens that are hardly traceable. Third, the ability for citizens to have a flawed image of the role of AS in the benefits system related to complexity of AS, combined with limited motivation and transparency, and lacking knowledge development and retainment by supervisors. Semi-structured interviews with experts in the field were done, and provided additional insight into governance of AS, as well as possible governance improvements within organisations and the broader supervisory landscape. The findings of interviews helped build the System-Theoretical Process Analysis.

Special attention went out to combining AA and STPA. AA provided a useful problem sketch and information for STPA. Methodological differences between the two analyses were used to create synergy: STPA included more actors and other components, and better showed the sociotechnical context. Where AA is typically inter-organisational and STPA intra-organisational, a SCS showing inter-organisational relations was a new addition to STPA, bridging the gap between the analyses. AA and STPA differ in how they treat values, where AA allows for multiple values to be included in a more nuanced manner than STPA. AA can show more context when it comes to values, while STPA risks a negative bias. The nuance AA offered to values was not explicitly used in STPA in this research. AA also provided input for the interviews, which in turn informed and validated both AA and STPA. This resulted in a synergy in how the methods can be used together.

The main contributions of this research are the proposal it makes for a sociotechnical description of AS and their harms, introducing concepts of STPA to describe AS and their harms: hazards, losses, constraints, understood through Safety Control Structures and Causal Scenarios. Such descriptions can be used to build shared understanding of AS and their harms. It showed how this can be constructed using an Actor Analysis, interviews, and STPA, describing how these methods work together. Moreover, the use of Actor Analysis in sequence with STPA showed how an inter-organisational view can be included in STPA, and pointed towards the possibility to add further nuance by using value descriptions discovered through AA in STPA. Recommendations were made to supervisory authorities and policy makers. Supervisors should work on central and structural knowledge creation to strengthen supervision, work on actionable transparency, and have eye for different types of AS. Policy makers should ensure supervisors have sufficient means, also in the future, work on meaningful solutions, and on the longer term consider how to structure SUWI and the role of AS therein. For all these recommendations, sociotechnical mappings are of importance. Future research can build on the proposal for a description of AS made in this research, and work with practitioners to increase empirical underpinnings and come to shared understandings. It can also work towards further including different value (conceptions) in STPA, for which this research can be a starting point.

Preface

Before you lies the outcome of not just the past months, but also that of six years at TPM. In my thesis, I used methods and skills that were taught to me in my first year here, as well as those taught to me in the in the final time as a student in Delft. I want to thank professors I have had the honour of working with in my time here, for the personal conversations I have had as a student and TA, which inspired me to choose my own path throughout CoSEM. I want to thank Tineke for dragging another Segeren towards Water Management. Okay, perhaps a specialisation in Ethics, Water Management, and IT is a bit rare, I am happy to have constructed my MSc the way I wanted. Besides, the more you know, the merrier, right? I greatly enjoyed researching Crayfish one week, and Machine Learning the next. Perhaps a true last show of great cooperations with those teaching at TPM was this thesis. I want to thank Roel and Haiko for their help in constructing this thesis, and their consideration of the time and space I needed throughout the time I worked on it.

I especially want to thank Roel for asking me to write this thesis in the first place. I have enjoyed the weekly chats, meetings, and other events. As a true sign of his enthusiasm about the topic, I left most meetings with more ideas than I arrived with. I greatly enjoyed picking my favourites and running with them.

I want to thank Gerald and Stefan for offering me a position as an intern, and their guidance as one. My time at AP flew by, and I've seen more than I can remember. It was impressive to see how new things came to be. Professionally and personally, it has truly been a welcome stay.

The past years have been challenging in many respects, but the past months were perhaps the most challenging of all. I want to thank my friends, who provided distraction or a listening ear, my parents, brothers, sister. I want to thank the friends I made in Delft. After all those hours in study rooms and next to the coffee machine, TPM has become a bit empty (time to leave!). When I arrived in my first year, I was told TPM-student are like Swiss knives. I am not sure if they still tell students this, but now that I am leaving, I feel more like a whole cutlery drawer.

Last, but not least, I want to thank my grandpa for a final thumbs up. Your name is on here, too.

Waar ik wegga, laat ik werelden na.

Wybe Segeren
Delft, Juni 2024

Contents

List of figures

List of tables

List of acronyms

ADM	Automated Decision-Making
AI	Artificial Intelligence
AP	Autoriteit Persoonsgegevens
ARK	Netherlands Court of Audit (Algemene Rekenkamer)
AS	Algorithmic System
BZK	Ministry of the Interior and Digitalisation (Binnenlandse Zaken en Digitalisering)
CoSEM	Complex Systems Engineering and Management
CRM	Netherlands Institute of Human Rights (College voor de Rechten van de Mens)
DCA	Department for the Coordination of Algorithmic Oversight (Directie Coördinatie Algoritmes)
ML	Machine Learning
SCS	Safety Control Structure
STAMP	System-Theoretical Accident Model and Processes
STPA	System-Theoretic Process Analysis
SUWI	Structure Executing Agency Work and Income (Structuur Uitvoeringsorganisatie Werk en Inkomen)
SVB	Social Insurance Bank (Sociale Verzekeringsbank)
SZW	Ministry of Social Affairs and Employment (Sociale Zaken en Werkgelegenheid)
UWV	Employee Insurance Agency (Uitvoeringsinstituut Werknemersverzekeringen)
ZBO	Autonomous administrative authority (Zelfstandig Bestuursorgaan)

Definitions

Safety	Freedom from unacceptable losses (Leveson, 2012).
Accident	Unplanned and undesired event that results in a loss (Leveson, 2012).
Hazard	A system state or set of conditions that, together with a particular set of worst-case environmental conditions, will lead to an accident (Leveson, 2012).
Algorithm	Process or set of rules to be followed to perform a calculation (Balayn & Gürses, 2021).
Algorithmic system	Digital system that employs algorithm(s).
Artificial Intelligence	Any technique that makes use of algorithms and enables computers to mimic human intelligence (Prins et al., 2021; Raub, 2018).
Machine Learning	Subset of AI that applies statistical techniques to improve at tasks with experience (Raub, 2018).
ADM-system	Tool that leverages an algorithmic process to arrive at some form of decision which may then drive further system action and behaviour (could exhibit AI insofar as they contribute to decision-making tasks normally undertaken by humans) (Diakopoulos, 2020).

1. Introduction

This chapter introduces the problem in focus, research questions and methods used.

1.1 Problem introduction

1.1.1 Use of algorithmic systems by executing agencies

In recent years, algorithms and Artificial Intelligence (AI) have become buzzwords. The frequently discussed technologies have taken their place in the world around us (Diakopoulos, 2020; Dobbe, 2022; European Union Agency for Fundamental Rights (FRA), 2022). They potentially have great benefits for society, including increased efficiency and innovation, and are used by a plethora of private and public organisations (FRA, 2022; Prins et al., 2021; Raub, 2018). In the Netherlands, public organisations tasked with execution of policy (*executing agencies*) have sought to reap these benefits. In the last decades, these executing agencies have digitalised. Automated systems now perform part of the work the agencies are tasked with, in order to more efficiently perform their public tasks and to provide services more quickly and easily (Hamer et al., 2022; Tijdelijke Commissie Uitvoeringsorganisaties, 2021). Laws that in the past were used by professionals to make decisions have been translated into decision rules that are used in these automated systems (Widlak, 2022), whose efficiency represents great public value (Versmissen & Soerjadi, 2022). Executing agencies have become 'decision making factories', that use automated decision making at scale, and are interconnected by exchange of data (Stichting Kafkabrigade, 2023). Standardisation and digitalisation are said to not only increase efficiency, but possibly also neutrality of decisions (Tijdelijke Commissie Uitvoeringsorganisaties, 2021; Zouridis et al., 2019).

1.1.2 Negative impacts

However, while these systems can have great positive impact, so much can also be said for their possibly negative impacts (Raub, 2018). The new role of algorithmic systems in important services and infrastructures has led to the emergence of new risks for society (Autoriteit Persoonsgegevens, 2022; Dobbe, 2022). Furthermore, while seen as neutral, these systems can in fact perpetuate biases and exacerbate discrimination (Balayn & Gürses, 2021; FRA, 2022; Raub, 2018). When it comes to executing agencies, the move towards standardization and digitalization has had unintended consequences. Algorithms are more and more important for functioning of government, but citizens are not always the main focus when it comes to the use of these systems (Algemene Rekenkamer, 2021a). A standardised and digitalised approach might promise efficiency and neutrality, but in practice complicates finding solutions fit to individual cases (Tijdelijke Commissie Uitvoeringsorganisaties, 2021). Peeters and Widlak (2023) similarly note that, while digital government eases burdens for many, it can increase burdens in complex cases. Furthermore, the authors pose that as digital government becomes increasingly infrastructural, risks of administrative exclusion and increased administrative burdens might grow, and discretionary space for professionals (known as *'street-level bureaucrats'*) might further decrease. The weighing of interests in a decision has moved from professionals to automated systems, which means discretionary space has moved from professionals to developers (Zouridis et al., 2019). The shift from professionals using laws to make decisions towards translation of these laws into algorithmic systems influences weighing of decisions as well as the applicability of checks and balances that were designed for government employees making decisions (Widlak, 2022; Zouridis et al., 2019). Moreover, the use of algorithmic systems in government poses challenges to principles such as legitimacy, accountability, and transparency of government (Grimmelikhuijsen & Meijer, 2022; Wieringa, 2023).

Examples of harmful algorithmic systems have surfaced in the Netherlands. In the well-known childcare benefits scandal, thousands of citizens were wrongfully accused of fraud, partly due to an automated risk assessment (Prins, 2021). In 2020, the anti-fraud system SyRI (System Risk Indication), was banned due to human rights violations (Rechtbank Den Haag, 2020). Even before it was implemented, it had been criticized for lacking proportionality and subsidiarity (Wieringa, 2023). Furthermore, in the last two years, it was reported a contested algorithmic system guiding Visa applications made use of profiling, possibly leading to discriminatory outcomes (Rengers et al., 2023), a system used to combat fraud amongst students was found to disproportionately flag students with migration backgrounds (Ersoy & van der Gaag, 2023; Vié, 2023), and the Dutch Immigration and Naturalisation Service was reported to have used nationality as a risk indicator in a data-driven system (Hijink, 2022).

1.1.3 Regulatory responses

In response to such cases, legislators have pushed for increased regulation of algorithmic systems (Rijksoverheid 2022; van Huffelen, 2022b). The Dutch national coalition agreement of 2021 underlined recognition of fundamental civil rights in the online space, and stated an algorithm supervisor should safeguard monitoring of algorithms on transparency, discrimination and arbitrariness (*Coalitieakkoord 2021-2025*). In 2022, Autoriteit Persoonsgegevens (AP) was appointed to house this National Algorithm Supervisor (van Huffelen, 2022a). The Dutch DPA was already part of AP, and the new Department for the Coordination of Algorithmic Oversight (DCA) of AP will help signal, share, and analyse risks between different sectors. Policy documents put emphasis on the importance of cooperation and coordination between different sectoral supervisory authorities, sharing of knowledge between them, and shared understanding of norms that come into play (Rijksoverheid 2022; van Huffelen, 2022a, 2023b). Following the increased regulation by the European Union, Dutch supervisory authorities have already sought cooperation in the digital domain, aware of the new challenges supervision over the digital domain bring about (Houtman, 2023; van Huffelen, 2023b). In the near future, the recent *EU AI-Act* aims to enforce further regulation on AI-systems, and requires a national supervisory authority (AP, 2022; European Commission, 2021). AP poses that it makes sense for them to fulfil this role, as they already house the National Algorithm Supervisor (AP, 2022). It is thus clear there are ongoing efforts to organise regulation of algorithms in the Netherlands, with the DCA having a central role, and with attention to working together with different parties across sectoral bounds.

Aside from these formal decisions, policy and law, there have been other initiatives. Over the past few years, a plethora of guidelines, assessment methods and frameworks have emerged. For example, the *algorithm assessment framework* by the Netherlands Court of Audit helps government agencies assess whether algorithms adhere to quality criteria, and if the organisation has an overview of risks (Algemene Rekenkamer, n.d.-b). The *ethics guidelines for trustworthy AI* by the High-Level Expert Group on AI highlights what is required of AI-systems in order to be 'trustworthy', and includes a self-assessment list (High-Level Expert Group on AI, 2020). The *guidebook non-discrimination by design*, made for the Ministry of the Interior, explains which concepts are important to prevent discrimination by AI-systems, and does so from technical, legal, and organisational viewpoints (van der Sloot et al., 2021). The *assessment framework* by the Netherlands Institute for Human Rights (CRM) also aims to prevent discrimination, and focusses on helping government officials assess whether risk profiles will lead to discrimination in a legal sense (College voor de Rechten van de Mens, 2021). A more comprehensive assessment for human rights is the *impact assessment human rights and algorithms (IAMA)*, which consists of questions that need to be discussed when government

agencies aim to use algorithms, in order to signal possible risks on time (Gerards et al., 2021). The concept *implementation guideline responsible use of algorithms* by the Ministry of the Interior draws on all aforementioned works and aims to give government agencies guidance for responsible use of algorithms (Ministerie van Binnenlandse Zaken en Koninkrijksrelaties, 2023b). The Ministry of the Interior is currently working on a more general Algorithm Framework (Digitale Overheid, 2023). Furthermore, in 2022, an *algorithm register* for the Dutch government was published, which aims to improve transparency of algorithm use by government, and thereby also improve trust in government (Ministerie van Binnenlandse Zaken en Koninkrijksrelaties, 2023a). As of late 2023, it held more than 250 algorithms used by government (Overheid.nl, n.d.-a).

1.1.4 Complex Socio-technical Systems Thinking

Responses by regulators can be too narrowly focussed on technocentric solutions, while algorithmic systems and their impacts are not purely technical or social (Balayn & Gürses, 2021). Diakopoulos (2020), for instance, noted that Automated Decision-Making (ADM) systems are complex socio-technical systems comprising of human and nonhuman components. Bauer and Herder (2009) state that systems can be considered socio-technical if 'technological components and social arrangements are so intertwined that their design requires the joint optimization of technological and social variables'. The social and technical sides of such systems are intertwined and co-evolve, but one side never fully determines the other. Wieringa (2020) describes the presence of this intertwinement in algorithmic systems, and defines them as socio-technical systems that can be 'viewed, used, and approached from different perspectives'. Socio-technical systems can be considered complex. Complex systems are unpredictable and behave in non-linear ways: it is not always possible to predict outcomes when starting conditions are known, in contrast to complicated systems, which behave in patterned ways and are predictable (Sargut & McGrath, 2011). A simple algorithm by itself, for instance, might be complicated – it is when it is used by actors in the real world when complexity shows up. The dynamics of elements making up complex systems and the interactions with the outside world cause emergent behaviour (Bauer & Herder, 2009; Brazier et al., 2018). This means that overall system behaviour is more than a sum of behaviour of elements (Kamensky, 2011). For decision makers, having an overview of an entire complex system is difficult, described by Sargut and McGrath (2011) as a 'vantage point problem'. According to the authors, managers face unintended consequences and have difficulty making sense of situations. Managers have trouble forecasting the future, mitigating possible risks, and making trade-offs. Trade-offs, however, are common in complex socio-technical systems, as different design goals and values come together that might be complementary but can also conflict. This can also be seen at executing agencies, who for instance want to work efficiently, but at the same time want to be careful and hold the human dimension into account (Tijdelijke Commissie Uitvoeringsorganisaties, 2021).

One-sided views, such as technocratic solutions, are not sufficient in complex socio-technical systems. Similarly, viewpoints that simply point to human operators fail to comprehend the multiplicity of causes of accidents common to complex systems (Elish, 2019). Algorithms cannot be viewed in isolation, and the governance of systems such as AI requires a broader, systems perspective (Dobbe, 2022). According to Hamer et al. (2022), algorithms implemented within organisations, such as the algorithms in use by executing agencies, shape an 'algorithmic practice'. This includes not only the algorithm, but also factors such as expertise, organisational structures, technologies, and so on. If this algorithmic practice as a whole lacks sufficient safeguards, this can cause risk. Similarly, in an article about use of algorithms by

government, Grimmelikhuijsen and Meijer (2020) state that a perspective broader than just the technical dimensions is needed to shape responsible use of algorithms in organisations. Looking at an algorithm as a standalone entity is not sufficient, as it only gains meaning when put into the context of its use. These broader views of looking at safe use of algorithmic systems map to Systems Safety thinking, in particular as described by Leveson (2012), whose framework for safety of complex socio-technical systems is based in systems thinking (Aven, 2022).

1.1.5 Systems thinking and Systems Safety

Systems thinking is a way of analysing and designing socio-technical systems (Bauer & Herder, 2009). It is interdisciplinary and analyses systems as a whole, seeing them as 'a combination of interacting elements organized to achieve one or more stated purposes' (Faulconbridge & Ryan, 2014). In the case of complex socio-technical systems, those elements can be both social and technical in nature. In her work, Leveson drew lessons from systems thinking and developed a new way of thinking about system safety, bringing about new concepts of safety fitting to complex socio-technical systems. Leveson rejects older methods of analysing safety, that focussed on dividing systems into components and ensuring as little as possible went wrong by improving component reliability and controlling varying conditions of systems (Aven, 2022; Leveson, 2020; Leveson & Thomas, 2018). Instead, safety requires designing the entire socio-technical system so that it remains safe under varying conditions. In complex socio-technical systems, system decomposition and a focus on component reliability does not always ensure safety, as emergent behaviour does not always allow for controlling system behaviour through system decomposition alone (Dobbe, 2022; Leveson & Thomas, 2018). Methods described by Leveson allow to analyse systems as a whole and include both social and technical components of socio-technical systems (Leveson & Thomas, 2018).

According to Leveson (2012), a new view on safety was needed in part due to new types of accidents stemming from digital systems and software. Since algorithmic systems are complex socio-technical systems, this new view on systems safety can be a valuable framework to prevent accidents, such as mentioned cases that were already seen in the Netherlands. Leveson introduces the terms accidents, losses and hazards to describe safety of systems. In Systems Safety, the negative impacts on citizens and society in these cases can be considered *losses*, the cases themselves *accidents*. An accident, in turn, is the result of a *hazard*, which is 'a system state or set of conditions that, together with a particular set of worst-case environmental conditions, will lead to an accident' (Leveson, 2012). Leveson noted that current approaches to risk assessment, hazard analysis and accident analysis do not sufficiently incorporate management, social factors, and organizational factors, which are of importance to complex socio-technical systems. Cases used in this book on systems safety draw from sectors such as aerospace, commercial aviation and defence. This is not remarkable, as systems theory and systems safety thinking have their roots in defence and space systems (Leveson, 2020). It does however, point to a lack of application in other domains. For instance, Dobbe (2022) underlines the need for a broader systems perspective on safety and AI, and applies lessons from the work of Leveson to AI-systems, a domain where systems safety is yet to be broadly applied. Leveson herself also noted that the System Safety approach can advance all industries, not just aerospace and defence (Leveson, 2020).

1.2 Research Positioning

The execution of laws in the social domain, as well as the enforcement thereof, has quickly digitalised using algorithmic systems. In this rapidly developing field, not only do technological developments happen at speed, so have cases of harms by these technological advancements, followed by an institutional response with varying frameworks, reports, policy, the appointment of a Minister of digitalisation and the investiture of DCA as national algorithm coordinator. This brings forward the question of how these developments lead to safety or a lack thereof. A sociotechnical analysis can provide an overview of different actors and actions, describe how these can be valued in relation to safety or the potential for harms, and investigate what is still needed to govern algorithmic harms.

More authors have pointed to the need for sociotechnical analyses for creating a broader understanding of algorithmic systems in the public domain. Grimmelikhuijsen and Meijer (2020), in writing about responsible use of algorithms in government, pointed to the need for scientific literature that goes beyond the dichotomy of technical possibilities versus ethical risks. There should be attention for the way algorithms are embedded in governmental agencies and how they are used in practice. A plethora of values should be held into account not just during development, but also in subsequent stages. Zouridis et al. (2019), in an article on the shift of discretionary space in public bureaucracies, noted that in order to assess how discretionary space changes, research is needed into interactions between 'data scientists, software engineers, and the overall management as well as the politico-administrative relations that result from the new technologies'. In general, more work is to be done to understand the governance and use of algorithmic systems in the Dutch social domain. Although goals of digitalisation of government are known, less is known about possible side-effects (Widlak, 2022). Questions also remain when it comes to supervision of algorithmic systems in the social domain. Firstly, how those politically responsible supervise algorithms and data in the social domain (van Atteveldt et al., 2024). And secondly, little is known on how executing agencies organise safeguards around algorithmic systems, especially when it comes to human rights' issues (Hamer et al., 2022).

Thus, where developments on use, harms, and governance of algorithmic systems in the social domain rapidly arise, there is still too little socio-technical insight into these aspects. A Systems Safety Analysis, as a sociotechnical analysis with a focus on safety, can be of use to analyse algorithmic systems in the social domain. In recent work, Dobbe (2022) has pointed to the lessons that can be drawn from Systems Safety for ensuring safety of algorithmic systems. A recent article by Delfos et al. ([Under Review]) identified risk factors in the use of Machine Learning (ML) systems. It concluded that risk factors in view of professionals in the field include both technical and non-technical factors, but also that the field of System Safety could provide additional, important risk factors. There is thus still more to learn from Systems Safety when it comes to the use of algorithmic systems in the public domain.

This research aims to bring additional insights into the use of algorithmic systems by executing agencies in the social domain. It will do so by using socio-technical analyses, as is critical to understanding algorithmic systems, as they are socio-technical systems. In particular, it will leverage a Systems Safety approach to describe the governance of algorithmic systems both within executing organisations as well as in the wider landscape in which policy actions have taken place. This allows to understand how the potential for hazards remains, thereby giving the ability to inform supervisors and policy makers on additional interventions that might be needed in the future.

1.3 Research scope

The scope of this research will be the use of algorithmic systems by executing agencies that execute social security policies under the so-called SUWI-law. Under this law, UWV most importantly executes employee insurances, such as unemployment benefits, sickness benefits, and disability benefits (Houtzager & Verbeek, 2022). SVB executes social insurance policies, such as child benefits, general pension, and benefits for relatives (Houtzager & Verbeek, 2022). Both for the execution of these policies and enforcement of surrounding rules, the executing agencies make use of algorithmic systems. The use of standardisation and digitalisation by these executing agencies has a longer history and can be seen in relation to political developments. Since the nineties, new public management (NPM) led to increased focus on efficiency of government, which together with increased technical possibilities led to the use of algorithmic systems and data exchanges between governmental organisations (Houtzager & Verbeek, 2022; van Atteveldt et al., 2024). Particular attention has gone out to risk profiles: digital advancements together with a political focus on fraud led to increased use of data and algorithmic systems for enforcement of policies (Houtzager & Verbeek, 2022; Olsthoorn, 2016; van Atteveldt et al., 2024). However, the executing agencies also make use of automated decision-making, with its own benefits and risks (Widlak, 2022). The executing agencies in the SUWI-law are thus a good example of the coming together of social and technical components, for which a socio-technical analysis is key.

This use of algorithmic systems by UWV and SVB poses challenges for governance: they use algorithms to efficiently execute policy, but also need to ensure these algorithms don't impose harm. Both sides of this equation can cause harm to individuals and society. Kuziemski and Misuraca (2020) describe this as a 'tragic double bind' that the public sector is in, as they need 'to govern algorithms, while governing by algorithms'. DCA as external supervisor faces a similar challenge. Supervision is necessary as accidents lead to significant losses, but more rigid forms of supervision (e.g. blocking use of AS, fines) might similarly impact society. Although policy instruments have been introduced (see paragraph 1.1.3), there is little overview of how the algorithmic systems they target are governed. Furthermore, it is unclear how negative consequences arise from the use of the systems, and how the policy instruments aim to prevent these consequences. This work therefore focusses on creating a description of algorithmic systems in use at the mentioned executing agencies, and introduces hazards in this description in order to address possible negative consequences of use of these systems.

1.4 Research goal and research questions

The goal of this research is to show how algorithmic systems in use in the execution of social security schemes, as well as their possible hazards and the governance actions surrounding these systems, can be described using a sociotechnical systems lens. This description aims to build knowledge on the systems and their governance, as well as a shared understanding of the systems and their possible harms. This description can be used to inform governance actions, informing policy makers and supervisors such as the DCA. This research thus aims to answer the following main research question:

How can a systems safety perspective combined with an actor perspective be used to provide a sociotechnical description of algorithmic systems, their governance and possible hazards at agencies executing the SUWI-law?

1.4.1 Sub-questions

In order to answer the main research question, several sub-questions have been used, focussing on methods used and theoretical and practical understanding of the governance of algorithmic hazards. The first sub-question focuses on the actor field surrounding algorithmic systems and is answered using an Actor Analysis:

How can an Actor Analysis be used to describe the actor field surrounding algorithmic systems at agencies executing the SUWI-law?

The second sub-question is used to describe AS and their potential hazards, and uses a System-Theoretical Process Analysis:

How can a Systems Safety Analysis be used to describe algorithmic systems at agencies executing the SUWI-law, and the potential hazards of these systems?

The third sub-question builds empirical insights using semi-structured interviews. These interviews are used to better understand potential for hazards and possible governance improvements, but also give additional input for the two mentioned analyses and their underlying assumptions.

In practice, what can be learned about the governance of algorithmic hazards at executing agencies, and how can this be understood in the context of an Actor Analysis and Systems Safety Analysis?

The fourth and last sub-question focusses on the use of Actor Analysis and a System-Theoretical Process Analysis (STPA) together to provide a description of AS:

What are similarities and differences between Actor Analysis and System-Theoretical Process Analysis, and what do these mean for combined usage of these methods in describing AS?

1.5 Alignment with CoSEM-programme

This research is done as a capstone project of the Masters programme Complex Systems Engineering and Management (CoSEM). This master's programme aims to teach students to analyse and design in complex socio-technical systems, where attention for the multi-actor and multidisciplinary nature of problems and solutions is of great importance. The topic of this research falls under the focus of CoSEM, in particular the specialisation in IT. As explained in the introduction, algorithmic systems can be seen as complex socio-technical systems. The use of these systems is a coming together of technical artefacts with a social environment and institutional artefacts. Algorithmic systems are governed by humans and law, and are used to provide services to society. Not only the scope relates to CoSEM, so do the methods used. Structured analyses of both social and technical components of complex systems are at the basis of CoSEM and the faculty of Technology, Policy and Management. The holistic view that STPA and Actor Analysis offer, aligns with the way systems are analysed and designed for in CoSEM. This perspective is of value for this topic, as purely technical, social, or managerial lenses fall short in analysing in and designing for the context in which algorithms are nowadays employed.

1.6 Research outline

This report is structured as follows: after this introduction, Chapter 2 explains the methods that are used in this research project. Chapter 3 provides a description of the use of Actor Analysis and STPA in combination, as well as the basis for comparison of these methods. Chapter 4 describes the Actor Analysis, after which Chapter 5 describes the System-Theoretical Process Analysis. This is followed by Chapter 6, which explains how interviews were elicited, and discusses the findings thereof. Chapter 7 is a reflection on the combination of AA and STPA, relating back to Chapter 3. Chapter 8 includes the primary conclusions of this research, a discussion thereof, and important limitations and recommendations.

2. Methodology

This chapter provides a brief overview of methods used to answer the research questions. The general approach this research used is a qualitative one, using structured analyses based on insight from literature, reporting, documentation and interviews.

2.1 Actor Analysis

An Actor Analysis was used to describe the actor field surrounding AS and their governance. Actor analyses are of importance to complex socio-technical systems, as actors can rarely find and execute a solution to a problem on their own. In practice, multiple actors have to work together, with different ideas and interests coming together. In the context of the supervision of algorithmic harms, there are multiple supervisory bodies, governmental organisations and executing agencies that need to work together to prevent accidents and shape the system of governance of hazards. According to Enserink et al. (2022), actors can be persons, organisations, or social entities that are able to 'act on or exert influence on a decision'. Their behaviour, in short, can be described by the actor networks they are in, perceptions actors have of the world around them, the values they uphold, and the resources they have. An actor analysis, therefore, looks into these dimensions to better describe the problem situation, including different viewpoints on it.

In this report, the actor analysis served as a point of departure. By making a brief overview of the actors involved, their reason for involvement, their (formal) relations, their views and resources, a first context was sketched. The Actor Analysis thereby enriched the problem description, as is the goal of this type of analysis. The steps followed were that as described by Enserink et al. (2022), this approach is the basis of Chapter 4. The information needed for this analysis was found using a document analysis, whereby various types of documents were used to form a picture of the actors and the actor field. The overview created and knowledge gathered in this analysis informed the subsequent System-Theoretical Process Analysis (STPA) and the interviews that were elicited.

2.2 System-Theoretical Process Analysis (STPA)

To get a broader overview of the way algorithmic hazards are managed in this domain, a Systems Safety analysis was done. As mentioned, Leveson rejected methods that see dealing with safety as a matter of increasing reliability of system components, thereby preventing component failures (Aven, 2022; Leveson, 2020). Leveson (2012) put forward the analysis model *Systems-Theoretical Accident Model and Processes (STAMP)*, that focusses on safety as an emergent property of a system as a whole instead. Rather than ensuring safety through increasing component reliability, STAMP focusses on ensuring constraints on system components and their interactions, thereby ensuring the system as a whole remains safe. Safety under STAMP is the freedom of losses that are deemed unacceptable by stakeholders involved in the system. These losses stem from hazards, system states that together with worst-case environmental conditions will lead to losses. Constraints aim to prevent these hazards from occurring, thereby ensuring safety. Three elements are used in STAMP to describe sociotechnical systems. Firstly, Safety Control Structures (SCS) display components of systems hierarchically, as well as feedback and control actions between the components. The components are controlled processes, and the controllers that control these processes. For example, a controlled process can be the awarding of social benefits, executed by use of AS. Typical controllers are then the employees involved, as well as developers of AS and higher-up managers. Secondly, STAMP uses constraints on the components displayed in the SCS, and

the interactions between these components. These constraints are thus used to ensure safety of the overall system by preventing hazards. Thirdly, STAMP uses process models. Process models are models that operators have of the process that is to be controlled or of automation that is used.

Based on STAMP, a hazard analysis method was created, which was used in this research. *System-Theoretic Process Analysis (STPA)* builds on STAMP and uses the three elements: Safety Control Structures, constraints and process models. STPA was developed as a proactive method, used to investigate accidents before they take place (Leveson & Thomas, 2018). Asides from the three elements of STAMP, STPA includes more causal factors for accidents, among which are social, organisational and managerial factors. STPA was based on documents from the two executing agencies (UWV and SVB), reporting on AS in the social domain, interviews, and the outcomes of the Actor Analysis. A more extensive explanation of Systems Safety and STPA is included as an introduction under Chapter 5.

2.3 Semi-structured interviews

Interviews have been used to gain further insight into the governance of AS in the social domain. Semi-structured interviews were used to gain empirical insights and gather information that is not (yet) explicitly documented. Furthermore, interviews added to validity of analyses done in this research, and allowed to contextualise and actualise information. The Actor Analysis served as a basis to identify possible interviewees. Interviews were sought within the actor field, with the aim of interviewing professionals within the different organisations in this field. Semi-structured were used as they allow for flexibility and reciprocity between interviewer and interviewee, enabling exploration of the field of study (Kallio et al., 2016). In order to inform previous analyses, and provide additional information into governance of AS in the social domain, four global themes were used in the interviews:

1. What does the interviewed organisation do (role, means, actions)?
2. What does interviewee see in their respective role (observations on workings of governance)?
3. What are possible governance improvements within organisations using AS?
4. What are possible governance improvements within the wider supervisory field surrounding AS in the social domain?

The first theme was used to test assumptions of the Actor Analysis and STPA, and gain further insight into structure of the governance of AS and different roles of actors therein. The latter three themes were analysed by use of a thematic analysis, whereby parts of interview transcripts were thematically marked and flagged with descriptions, also called coding (see for instance Harding and Whitehead (2013) or Liamputtong (2008)). These gave the ability to further inform STPA, wherein observations could be used to understand possible losses, hazards, and the causal scenarios leading to them. The possible governance improvements were be used to build further and inform possible additional constraints and interventions.

3. Synthesis: Actor analysis and STPA

This chapter describes initial considerations regarding Actor Analysis and System-Theoretical Process Analysis. This serves as a basis for answering the following research question:

What are similarities and differences between Actor Analysis and System-Theoretical Process Analysis, and what do these mean for combined usage of these methods in describing AS?

These considerations form the basis of comparison of the two methods. This comparison can be found under Chapter 7.

3.1 Need for actor analysis

STPA offers a structured way to analyse safety systems surrounding the use of algorithmic systems. However, STPA offers little guidance on building a Safety Control Structure, although this is a central artefact in the analysis. When it comes to actors in this control structure, different publications involving STPA do not describe how actors can be identified, or how the actor field ties into the safety structure surrounding AS and the processes they are part of. Leveson (2012) and Leveson and Thomas (2018) describe that stakeholders should identify what is unacceptable to them, i.e. the losses they think should be prevented. Other than this, the authors do not provide tools to analyse the actor field. Actor Analysis can fill this gap in STPA, and increase focus on actors and the normative and political considerations within the actor field. Complex socio-technical systems often comprise a plethora of actors that are of importance in solving problems. Enserink et al. (2022) mentions that where technical solutions are generally sufficient to solve complex technical problems, in socio-technical systems 'the complexity of players caused by different perspectives and conflicting interests implies that solutions either need to be negotiated or are imposed by a party with the power to do so'. This multi-actor dimension of sociotechnical problems can be a valuable addition to STPA. AA can help build a SCS and increase focus on actors involved in the system. An exploration of actors involved in the problem situation through an actor analysis can serve as a vantage point through which the problem is explored.

Not only does AA fit in combination with STPA, the converse is also true. STPA fits in with the context in which an actor analysis is usually done. Enserink et al. (2022) describe Actor Analysis as part of problem synthesis, turning an initial problem perception into a rich problem perception. In this synthesis, there are three analyses: an actor/network analysis, a causal/system analysis, and a future scenario analysis. These are not necessarily done in sequence, but can be done together and iteratively, strengthening each other. STPA covers two of these analyses specifically for safety systems: the causal/system analysis and the future scenario analysis, but not the actor/network analysis. A safety control structure is a system analysis, and a subsequent step in STPA is identifying causal scenarios. STPA also looks into degradation of controls over time, which can be seen as a form of a future scenario analysis. In this way, STPA fits together with an actor analysis, as is displayed in Figure 1.

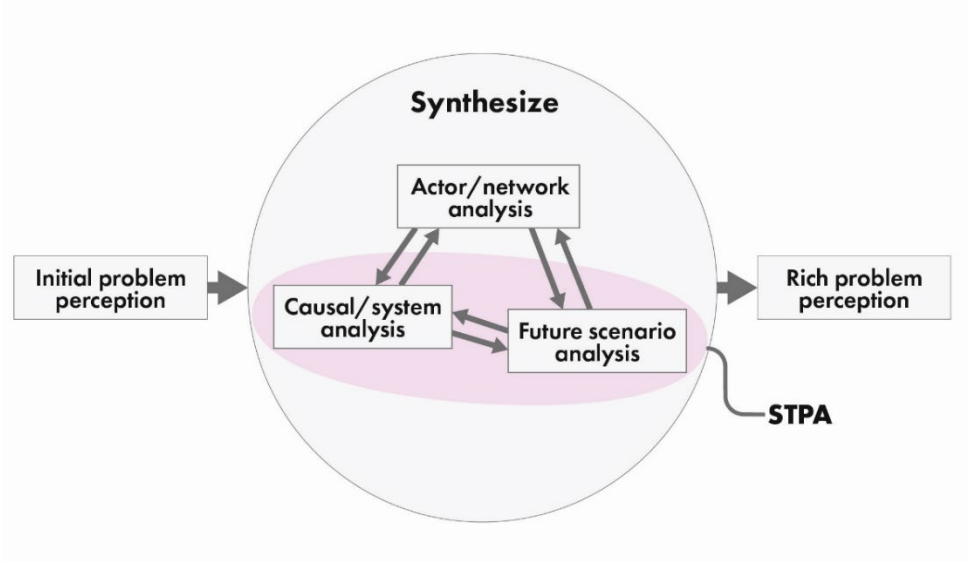

Figure 1 STPA in the context of problem analysis (based on Enserink et al. (2022))

3.2 Comparison of Actor Analysis and STPA

To compare the methods, their purpose, the methods used, and the role values play will be discussed. A summary of the comparison of these aspects can be found in Table 1. The influence of these differences on outcomes of the analyses will be discussed in Chapter 7.

3.2.1 Purpose and application

The two analyses are executed for different purposes. Framed around a problem situation, actor analysis (AA) is used for problem analysis, whereas STPA is used to identify problems. AA works from an initial problem statement with a multi-actor component and works to enrich it. STPA, on the other hand, is used to identify the potential for accidents to occur. The specific problem is not necessarily known a-priori, although the potential for a safety problem is assumed. STPA then works to identify what this safety problem specifically is.

3.2.2 Methodological differences

AA and STPA differ methodologically in their scope and how they are structured. When it comes to scope, AA typically looks at different organisations and their overarching characteristics, as well as formal relations between organisations. STPA typically looks within one or few organisations, describing processes within them. The unit of analysis under AA is actors, whereas STPA focusses on controlled processes and controllers, including systems, organisations, and individuals. Here, the difference in scope is also reflected. Under AA, actors are typically organisations or formal representation of individuals. STPA does not uphold such a strict notion of actors. Both methods place the units of analysis in a hierarchical structure, under AA this is the formal chart and under STPA the safety control structure. The relation between the units differs, under AA the formal chart describes formal relations between actors. STPA describes control actions and feedback between controllers and processes.

3.2.3. Values

Lastly, the two methods differ significantly in how they treat values. Actor Analysis includes values insofar as they are relevant to actors. The conception of these values depends on those of actors, as does the relation between different values. AA has a contextual and descriptive view on values, including safety. STPA, on the other hand, includes values insofar as they are relevant to safety. The conception it has of values is dependent on the influence they have on safety. The relation between values is hierarchical: if STPA takes into account values, safety is always the highest objective, and other values influence the ability of a system to reach this objective. Under AA, the different values could be understood in a different hierarchy, or even as equal, depending on the viewpoints of actors.

When it comes to safety in particular, STPA always describes this as the absence of losses. It thus prescribes what safety is, depending on the losses taken into account. AA looks at safety as a value that can be understood differently within the actor field. It thus describes how safety is understood within the actor field. Different actors might have different notions of safety, or differ in the extent to which safety needs to be upheld in order for a system to be considered safe. STPA is more definitive in its description of when a system is considered to be safe.

Table 1 Comparison of aspects of Actor Analysis and System-Theoretical Process Analysis

Aspect	Actor Analysis	System-Theoretical Process Analysis
Purpose		
Purpose of analysis	Problem analysis: enriching a known initial problem statement	Problem identification: potential for safety problem assumed
Methods		
Typical scope	Inter-organisational: looks at different organisations in problem situation	Intra-organisational: looks within one or few organisations surrounding relevant controlled process
Unit(s) of analysis	Actors: typically organisations or representation of individuals	Controllers and controlled processes (various)
Constellation of units	Hierarchy, formal relations	Hierarchy, control and feedback
Inclusion of units	Has capacity to influence decision-making or to act on decisions and their outcomes (actor)	Part of safety control structure: involved in the problem situation in practice (can control or is controlled)
Values		
Inclusion and relevancy of value(s)	Actor-induced, value plurality: different values relevant to actors in problem situation	Safety-related, limited value plurality: values relevant to safety
Conception of values and their relations	Conception of different values and their relations through conceptions of relevant actors	Conception of different values through their influence on safety, relations are hierarchical, where safety is the superior value
View on safety	Descriptive, contextual: dependent on actor	Prescriptive, definitive: absence of losses

4. Actor Analysis

This chapter describes the actors involved in the system of interest through an Actor Analysis. It makes use of the method described by Enserink et al. (2022), which is part of a problem analysis. The goal of this analysis is to answer the following research question:

How can an Actor Analysis be used to describe the actor field surrounding algorithmic systems at agencies executing the SUWI-law?

The Actor Analysis consists of six steps, displayed in Figure 2. What follows is each of the steps of the analysis.

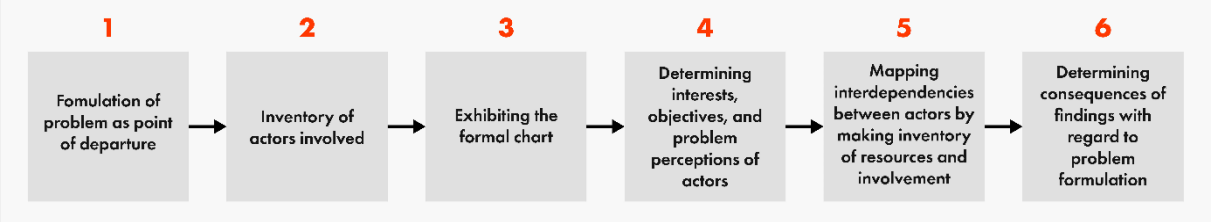

Figure 2 Steps of actor analysis

4.1 Problem as starting point

The problem situation follows from the introduction. In the execution of the SUWI-law by UWV and SVB, which encompasses social benefits surrounding work and income, algorithmic systems (AS) are used to judge eligibility, pay the benefits, and oversee mis- and abuse of the benefit schemes. Use of these systems is seen as efficient and necessary, but can lead to harms to individuals and society. The problem situation serves as a frame for the analysis.

4.2 Inventory of actors involved

For the described problem, a first inventory of actors is made that has been iteratively improved. What follows is a short description actors and their involvement in the case.

Social Insurance Bank (SVB)

SVB is autonomous administrative body (ZBO) responsible for executing social insurance schemes (Sociale Verzekeringsbank, n.d.-d). In executing this task, they ensure citizens are informed, they judge whether and to what extent citizens are entitled to certain social benefits, pay those who are entitled and enforce rules. They do this in a changing society, where expectations of citizens are changing due to digitalisation, and where political actors seek quick results (Sociale Verzekeringsbank, n.d.-b). SVB uses AS to execute their tasks, both in judging entitlement to policies and enforcement of these policies (Sociale Verzekeringsbank, n.d.-c). The organisation has internal normative frameworks and tools surrounding AS, developed by their ethics center (Sociale Verzekeringsbank, 2020). The ethics center also organises Intention Workshops (garages de Bedoeling), where issues in execution are discussed with a focus on the deeper meaning and goals of laws and policies. These workshops inform policy, tools, and include issues with AS (Berg et al., 2019).

Employee Insurance Agency (UWV)

UWV is a ZBO responsible for executing employee insurance policies, such as the unemployment insurance act (WW), Invalidity Insurance Acts (WAO, Wajong), and Sickness Benefits Act (Ziektewet) (Business.gov.nl, n.d.; Uitvoeringsinstituut Werknemersverzekeringen, 2023b). UWV also makes use of AS in execution of their tasks (uitvoeringsinstituut Werknemersverzekeringen, n.d.). Aware of possible risks, they see the use of AS and data-analysis as a necessary and effective tool (Uitvoeringsinstituut Werknemersverzekeringen, 2023b). To this end, they have developed several AS, including risk scans, decision tools, and ADM-systems (uitvoeringsinstituut Werknemersverzekeringen, n.d.). Internally, UWV has developed policy surrounding these AS, such as ethical guidelines, a risk scan policy, and installed an ethical committee (Uitvoeringsinstituut Werknemersverzekeringen, 2021a, 2021c).

Ministry of Social Affairs and Employment (SZW)

Both agencies fall under the responsibility of the Ministry of Social Affairs and Employment (SZW), and execute its policies. SZW is responsible for oversight of the organisations, the way they execute policies, and their finances, although the ZBOs execute policy independently (Ministerie van Sociale Zaken en Werkgelegenheid, 2021). Even so, the three parties do work together to improve policy execution (Ministerie van Sociale Zaken en Werkgelegenheid, 2023b). This includes cooperation when it comes to development and use of AS (Uitvoeringsinstituut Werknemersverzekeringen, 2021b, 2022).

Autoriteit Persoonsgegevens (AP) and Directorate Coordination Algorithms (DCA)

When zooming in on oversight, more organisations come into view. Autoriteit Persoonsgegevens is the organisation that houses the Dutch Data Protection Authority (DPA), which oversees organisations when it comes to the GDPR. AP falls under the responsibility of the Ministry of Justice and Security, who also nominates members of their board ("Uitvoeringswet Algemene verordening gegevensbescherming," 2021, 1 July). In 2022, the State Minister for Digitalisation (part of the Ministry of the Interior and Kingdom Relations) appointed AP to house the national algorithm coordinator (van Huffelen, 2022a). The Department for the Coordination of Algorithmic Oversight (DCA) has taken up this task. The task of the DCA originally included signalling, analysing, reporting on algorithmic risks, strengthening cooperation between supervisors and bringing about guidance and shared understanding of norms and values (van Huffelen, 2022a). In this work, there was to be special attention for transparency, discrimination, and arbitrariness. DCA has since published two risk reports, discussing overarching developments, several specific cases, and relevant law and policy (Autoriteit Persoonsgegevens, 2023b, 2023c). Late 2023, the Ministry of Internal Affairs and DCA changed the task description of DCA to include looking into connection to international laws, policy, guidance, standards and norms, accessible information aimed at those who are at risk of being harmed by algorithms, and looking into contact points where citizens could notify of and get information about AS-related problems (van Huffelen, 2023a). Furthermore, the DCA follows ongoing developments surrounding the AI-Act and is helping to coordinate preparatory activities (Autoriteit Persoonsgegevens, 2024).

Ministry of the Interior and Kingdom Relations (BZK)

The Ministry of the Interior and Kingdom Relations (BZK) has a Minister for Digitalisation, who has a coordinating role for digital policy of the Dutch government (van Huffelen, 2022b). The appointment of DCA as national algorithm coordinator is part of the Value-Driven Digitalisation Agenda (Werkagenda Waardengedreven Digitaliseren), of which the aim is to ensure public values and human rights are protected, and to ensure a level playing field in the economy (Rijksoverheid 2022). Aside from the DCA, this agenda has other specific action points when it comes to algorithm regulation, such as the National Algorithm Register. A new action point the development of the so called 'Algoritmekader', which aims to help governmental agencies make use of AS in a lawful and ethical manner (Digitale Overheid, 2023).

Netherlands Institute for Human Rights (CRM)

The Netherlands Institute for Human Rights (CRM) aims to protect human rights in the Netherlands, and increase attention for and compliance with these rights, especially the right of equal treatment and non-discrimination ("Wet College voor de rechten van de mens," 2020, 1 January). The Institute can investigate human rights violations, judge and report, advise, and inform. One of the themes they focus on is digitalisation, noting that it has benefits for citizens and professionals, but can also impact the right to non-discrimination, freedom of speech and privacy, and that it cause exclusion and can make it harder for citizens to obtain justice (College voor de Rechten van de Mens, n.d.). The organisation investigates how government uses algorithms and how this can impact citizens. For instance, they advised the Senate on a law about data sharing between government organisations (coined the 'next SyRI'), and stated more safeguards were needed to prevent discrimination, specifically referring to automated data analysis (van Dooijeweert, 2021). CRM is also a ZBO and falls under the responsibility of the Ministry of Justice and Security, in coordination with BZK.

National Ombudsman

The National Ombudsman can investigate conduct of the National Government and governmental agencies (see art. 78a "Grondwet" 2023, 22 February). The Ombudsman is appointed by the House of Representatives. It is a place where citizens can file complaints against government, and where they are helped in resolving their situation, with an investigation as a possible outcome (Nationale Ombudsman, 2023). The Ombudsman is aware of the risks AS-use by government poses on citizens, and stated that government should work in a clear, open, and solution-oriented way in this regard (Govers et al., 2021). Ombudsman is a High Council of State (Hoog College van Staat), as is the Court of Audit (Overheid.nl, n.d.-b).

Court of Audit (ARK)

The Netherlands Court of Audit (ARK) is responsible for monitoring spending by the national government, as well as effectiveness of policy (art. 76 "Grondwet 2023, 22 February; Algemene Rekenkamer (n.d.-a)). ARK checks whether the government spends sensibly, carefully and economically, and in doing so ensure public accountability of spending of public funds (Algemene Rekenkamer, n.d.-a). One of the areas of focus is the use of (self-learning) algorithms. The Court of Audit developed an assessment framework to check whether organisations are in control of AS risks, and has since tested whether AS in use live up to the standards brought forward in this framework, which led to criticism on some algorithms in ARKs reports for the House of Representatives (Algemene Rekenkamer, 2021b).

Non-Governmental Organisations: Amnesty, Bits of Freedom

Amnesty International is a non-governmental organisation that wants to ensure human rights are upheld. It does so by lobbying, researching, impacting policy, and creating awareness, also on the topic of technology in use by government and the impact this has on citizens (Amnesty International, 2023c). To this end, the organisation has published a report about the Childcare Benefits scandal, in which it was critical of the Dutch government and pushed for increased regulation of the use of AS by government (Amnesty International, 2021). More recently, the organisation stated in a report that recent policy actions, including the National Algorithm Coordinator, were not enough to prevent new scandals (Amnesty International, 2023b).

Bits of Freedom is a NGO focussing on freedom of the internet in the Netherlands, especially on communication freedom and privacy (Bits of Freedom, n.d.-b). It has in the past been critical of transparency and explainability of decisions made with complex AS, stating that government should not make decisions it is not able to explain (Zenger, 2017).

House of Representatives (Tweede Kamer)

The House of Representatives is the formal lawmaker in the Netherlands. It currently consists of fifteen parties representing citizens (Tweede Kamer der Staten-Generaal, n.d.). These parties come from different backgrounds and have different foci, which means they do not have a singular objective, goal or opinion. This of course reflects citizens themselves. In the party plans that were published in the last elections, various themes were mentioned frequently. Of the ten biggest parties in the House, a majority speak about a distrustful government, a need for transparency in AS-use or mandatory use of an algorithm register, and complexity or simplification of social security systems. Other important themes in the last elections were the human dimension, people's livelihoods, and the position of AP in supervision of AS (see appendix A.1).

The House of Representatives has several instruments it uses that are relevant to this case. The House is the formal lawmaker in the Netherlands and it has a role in controlling government. Instruments include lawmaking rights, ability to change laws, parliamentary inquiries, questioning ministers and secretaries of state, and so on (Tweede Kamer der Staten-Generaal, 2023). For instance, the house investigated and reported on the childcare benefits scandal. Furthermore, the house has a commission on digital affairs, in which members of the house discuss this theme.

National Client Council (LCR)

The SUWI-law states that executing agencies should work towards client participation. In the law, it is decided that UWV and SVB should organise client participation that allows clients to be involved in discussions, policy and goal forming, and give advice (see art. 7 "Wet structuur uitvoeringsorganisatie werk en inkomen" January 1, 2024). Furthermore, the law states that there should also be a National Client Council, wherein members of client participation of UWV and SVB are represented (see art. 8 "Wet structuur uitvoeringsorganisatie werk en inkomen" January 1, 2024). This National Client Council, the Landelijke Cliëntenraad (LCR), meets with UWV, SVB and the minister of SZW at least yearly.

In practice, these sections of the SUWI-law resulted in UWV and SVB each having their own client councils, who are represented in the National Client Council. LCR represents citizens that receive payment from the government, including those who receive social benefits as part of SUWI (National Client Council, n.d.). LCR consists of a network of organisations and sees

itself as a centre of knowledge about citizens, their needs and experiences, and ensures citizens are involved (National Client Council, 2023b). In this light, they advise lawmakers and executing agencies, add to the public debate and are part of several commissions active in this field (National Client Council, 2023a).

Part of the National Client Council are thus the client councils of SVB and UWV. In the client council of UWV there are only citizens who are 'clients' of UWV (that is, they receive aids from UWV) (Client Council UWV, n.d.-b). In their task, they signal where execution of law goes wrong, give solicited and unsolicited advice, make proposals for improvement, and can request information from UWV (Client Council UWV, n.d.-a). The client council of SVB not only has clients of SVB as members, but also representatives of advocacy organisations, such as unions and elderly associations (Sociale Verzekeringsbank, n.d.-a). Since LCR can be seen as a national overarching organisations that represents both client councils, in the analyses that follow, only LCR will be taken into account.

Other actors

Several actors are included and of importance indirectly, they will not be analysed further but are described briefly to provide context. On an international scale, the European Union is of importance. The European Union consists of several supranational institutes that influence this case mainly through the creation of laws and regulations. The two most frequently mentioned pieces of legislation in this case are the General Data Protection Regulation (GDPR) and the recently published Artificial Intelligence Act (AI-Act).

The Ministry of Justice and Security is responsible for the GDPR in the Netherlands, which is independently executed by the Dutch DPA. The ministry is the formal client of AP when it comes to the GDPR, and is thus responsible for providing budget for AP. This is also true for the DCA as it is part of AP, even though the formal client is currently the Ministry of Internal Affairs.

In the execution of laws, several governmental organisations exchange their data. These organisations include municipalities (personal records database), Netherlands Vehicle Authority (vehicle data), and the tax office (income and tax data). SVB and UWV also exchange data with each other. This is assumed to happen via SUWInet, which serves as a central data exchange (Schelfhout et al., 2022).

4.3 Mapping formal Institutions and Relations

The different actors in the playing field surrounding the problem at hand relate to each other. A formal chart describes formal relations between actors. As seen in Figure 3, it is organised with actors in order of hierarchy top to bottom. Actors can have a hierarchical relationship, a single-sided arrow pointing to the actor lower in hierarchy, or a representation or membership relationship, displayed by a two-sided arrow.

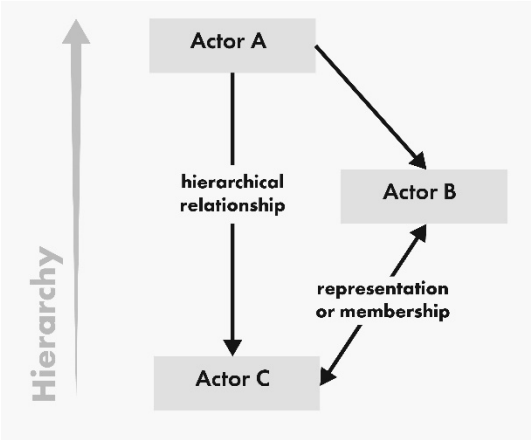

Figure 3 Example of relations in a formal chart

The formal chart of the case can be found under Figure 4. As can be seen, DCA is displayed as part of Autoriteit Persoonsgegevens, but only has a formal relationship with the ministry of the Interior. The NGOs have no formal relationships. This does not mean these parties necessarily have little influence on the problem at hand, but it does mean the power they have will more likely be informal, rather than formal power. The relation between LCR and the executing agencies is displayed as a two-sided arrow, since it is not immediately obvious which holds a higher hierarchy. That is, this is not clearly formally recorded, although it might be clear in practice.

The formal chart also displays several cooperations between actors. Several Government Inspectorates work together in the so-called Inspectorate Council (Inspectieraad) in order to improve their supervisory tasks (Rijksinspecties, n.d.). They organise the informal 'working group AI and Algorithms', which focusses on supervision of these technologies (Rijksinspecties, 2021). Over twenty organisations are part of this group, including AP, ARK, and CRM, gathering once every two months (Nas & Ouburg, 2022). Members are displayed with a blue dot. In 2021, AP, CRM, AFM and ACM started the 'Platform Digital Regulators' (SDT), aimed to investigate risks in a digitalising society, discuss new laws, regulations and overlapping issues (Autoriteit Consument en Markt, 2021). It has a group focussing on Algorithms and AI, for which the platform includes more regulators. SDT, together with other regulators, have been involved in discussions with policy meetings with BZK (see for instance: van Huffelen (2023b)). UWV and SVB work together with other executing agencies in the 'Manifestgroup', aiming to improve public service, also when it comes to digitalisation (Manifestgroep, 2013, 2022). It includes sixteen large executing agencies.

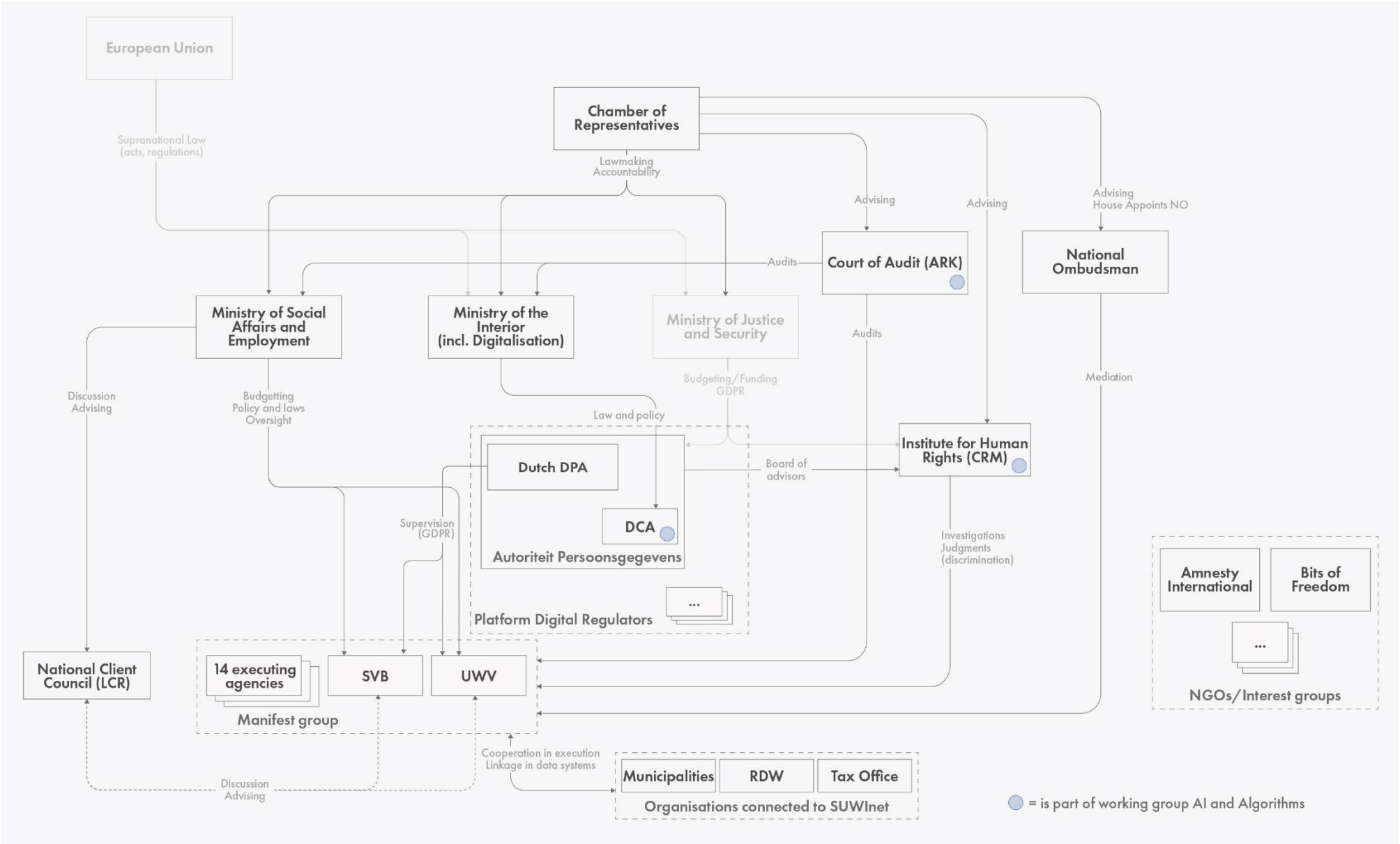

Figure 4 Formal chart of governance of AS in the SUWI-system

4.4 Identifying key actor characteristics

This next step focuses on the interests, objectives, perceptions, and resources of actors. The book by Enserink et al. (2022) prescribes a method using an overview table with the following categories;

- Actors;
- Their interests;
- Desired situation or objectives;
- Existing or expected situation and gap;
- Causes;
- Possible solutions.

Since the problem situation is ongoing, the possible solutions were replaced with current actions and means for each actor. This better reflects the fact that solutions are already underway, or even in place. It also focusses the analysis more towards the current safety system and thus makes it better fit to analyse gaps within this safety system. This version of the overview table is based on publications done by actors and thus describes their public stance and actions on the topic only. The overview table, where these factors are explained per actor, is included in appendix A.2.

4.4.1 Problem perceptions and actions within SUWI

For both UWV and SVB, major objectives are good public services, but also efficiency. They use AS in order to provide social security schemes efficiently to citizens, but also to enforce rules surrounding these schemes. Both organisations are aware that use of AS can cause risks to citizens and the organisations themselves. They take several actions to minimise these risks. Both organisations state the involvement of humans in the loop, have published algorithm registers, use advice of external organisations, and have procedures and roles surrounding privacy rights. In other ways, their approaches differ. Both organisations are active when it comes to ethical issues. UWV has an ethical committee, comprising of employees and external members, that can be asked for advice. They also have the internal guideline named the 'ethical compass', which outlines ethical baselines that should be followed. Furthermore, they have the 'Model Risk Management'-policy for risk scans. SVB, meanwhile, has an ethics centre which organises Intention Workshops (garages de bedoeling). Here, employees can discuss ethical issues. Based on these discussions, the ethics centre produces recommendations and policy.

4.4.2 Problem perceptions and actions in the broader landscape

In the broader landscape, there are obvious differences in perceptions and actions based on roles and goals of organisations. The different supervisory authorities and NGOs have specific foci, for instance on human rights or privacy issues. The ministries have a more overarching viewpoint. BZK has a focus on digitalisation, SZW on social security and abuse thereof. SZW, however, has a special role as it not only produces policy for execution of social security, it also needs to oversee whether UWV and SVB properly execute laws and policies.

Several important actions include the ability to make laws, pass judgements, and certain investigatory rights. Furthermore, there is a plethora of actions that do not have a direct formal influence but are part of the way in which organisations shape their work. These are published essays, advices, and reports, wherein opinions are stated. Additionally, there are several frameworks, each with a different focus, for instance by ARK and CRM.

DCA currently has no formal powers, therefore relying on informal ways of executing their task. This is currently mainly done through their bi-annual risk reports. It also means there is no supervisor focused solely on algorithmic harms that has formal powers. The different supervisory agencies that have formal powers do, however, have attention for digitalisation within their roles.

Apart from different ambivalences between actors, mostly relating to the formal role they have, there is also notably some ambivalence within different actors. That is because multiple goals need to be attained: there needs to be social security, but abuse must be fought against. Citizens must receive help they need, but it also needs to happen efficiently. These internal ambivalences are present most of all in the chamber of representatives, ministries, and thereby also in UWV and SVB, trickling down through the formal relations between them. The plurality of views within actors can be influenced through scandals that arise. This can lead to increased focus on the human dimension over efficiency, or increased focus of supervisory authorities on the issue of AS.

4.5 Summarizing interdependencies

This next step summarizes the resources actors have, followed by their criticality, dedication, and power and interest. This serves as a way to investigate who is affected by the problem, who has the power to do something, who wants to, and which actors might work together in solving the problem.

4.5.1 Actor resources

A first step is an overview of the resources actors have. This overview, found in Table 2, is a summary of current actions that were identified under the actor characteristics.

Table 2 Summarised actor resources

Actor	Important resources (summary of current actions)
EU	• Supranational law

House of Representatives	• Rights to shape laws and regulations
	  ○ Debates, discussions
	  ○ Right to propose motions
	  ○ Right to amend laws
	  ○ Right to propose laws
	• Rights to check government
	  ○ Budgeting rights
	  ○ Right to inquire
	  ○ Right of interpellation
	  ○ Right to propose motions

Court of Audit	• Investigations into government action • Reporting to parliament
Ombudsman	• Investigation of citizen complaints
CRM	• Investigations into human rights violations • Advising of government
SZW	• Law proposals and policy creation • Organisational oversight over UWV/SVB • Budgeting for UWV/SVB
BZK	• Law proposals and policy creation • Value-driven digitalisation agenda (incl. frameworks, register)
AP	• Preventative oversight • Repressive oversight • Norms and guidance shaping (w.r.t. personal data usage)
DCA	• Norms and guidance shaping (w.r.t. algorithms and AI) • Risk monitoring and signalling
SVB	• Execution of SUWI-law (using AS) • Enforcement of SUWI-law (using AS) • Ability to shape algorithm use and surrounding internal policy
UWV	• Execution of SUWI-law (using AS) • Enforcement of SUWI-law (using AS) • Ability to shape algorithm use and surrounding internal policy
LCR	• Consultation with UWV, SVB, SZW (in accordance with SUWI-law) • Information requests (UWV, SVB) • Signalling function (client perspective)
Amnesty International	• Lobbying and public opinion shaping (through cooperation, campaigns, investigations) • Legal actions
Bits of Freedom	• Lobbying and public opinion shaping (through cooperation, campaigns, investigations) • Legal actions

4.5.2 Actor alignment

The different actors can, based on their interests and means, be grouped based on their dedication, criticality, and alignment of objectives. Such an overview can be seen in Table 3.

Table 3 Actor alignment: dedication and objectives

	Dedicated actors		***Non-dedicated actors***	
	Critical actors	**Non-critical actors**	**Critical actors**	**Non-critical actors**
Similar or supportive interests and objectives	SVB UWV	DCA Interior and Digitalisation	Social Affairs	
Conflicting interests and objectives	Dutch DPA Human Rights Institute Court of Audit House of Representatives	Amnesty Bits of Freedom	Client Council Ombudsman	

The executing agencies (UWV and SVB) both face similar issues, and are at the core of the problem. Therefore, they are both dedicated and critical, and face similar objectives. When it comes to algorithmic systems, both parties undertake actions to limit harms but also face the need to use them in policy execution. Also dedicated and critical are the Dutch DPA, the human rights institute, the court of audit, the house of representatives, and citizens. The supervisory bodies (Dutch DPA, Human Rights Institute, Court of Audit) each have a specific focus on a supervisory task, mostly focussing on harms, and have an independent role, and should therefore be seen as conflicting. The house of representatives is hard to place. Even though they are important, different parties may have different foci (e.g. tackling misuse versus harms by systems), and different levels of dedication. These foci and levels of dedication can also change over time.

BZK and DCA pose requirements and can help to bring about better understanding, shared norms and values, but are not critical to push changes. They have reasonably similar objectives when it comes to the case, although somewhat more specifically focussed on algorithms and their possible harms than UWV and SVB, who are working within a broader system wherein algorithmic systems are just a part of their efforts, and where goals such as efficiency and scale are also important. Amnesty and bits of freedom, as non-governmental interest groups, can advocate for human rights. They are influential, but also not critical. Their specific focus on human rights makes that their view on the problem can be seen as conflicting with UWV and SVB, but the interests of the two parties themselves can very well align.

The ministry of social affairs, the national client council and the ombudsman can all be seen as critical. Social affairs has an important role in law, policy, oversight, and budget in the SUWI-system. Their interests most likely align with UWV and SVB, although have a broader point of view. However, very little could be found when it comes to their involvement in the issue of algorithmic systems, and therefore they should be seen as non-dedicated. Similarly, the client council and ombudsman seem to not be very involved in the oversight over algorithmic systems specifically. Their focus on citizens in particular makes that they can be seen to have somewhat conflicting interests to UWV and SVB.

4.5.4 Power-Interest grid

A power-interest grid places actors on two axes, relating to the power they have to make changes, and the interest they have in the case. This PI-grid can be seen in Figure 5.

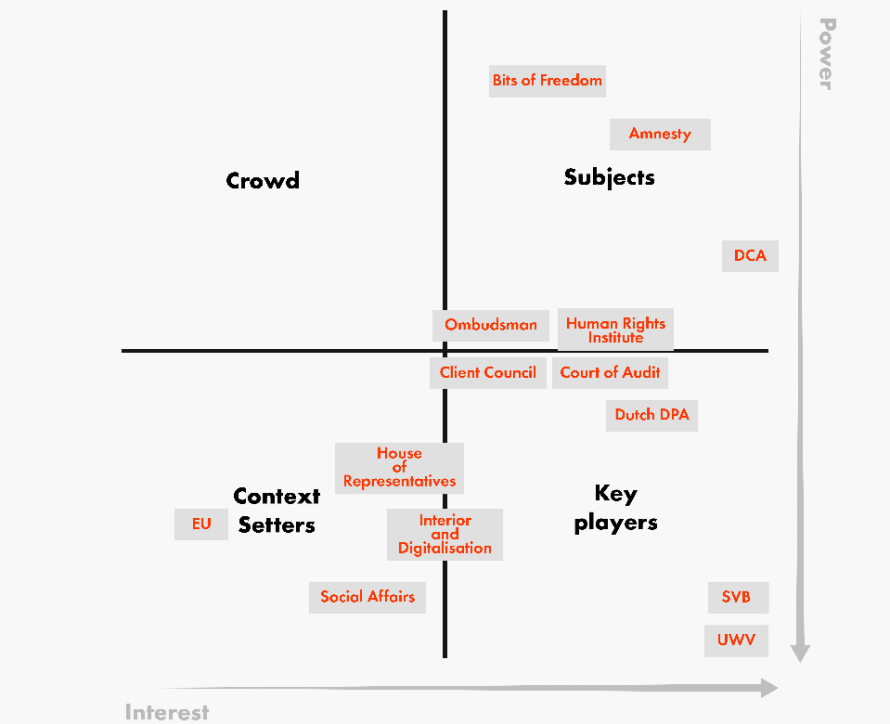

Figure 5 Power-Interest Grid

Roughly four groups can be identified in this PI-grid. First, the ministries, EU and House of Representatives. They are context setters. They have power in their ability to set laws, regulations, and budgets. However, they are not as interested as for instance UWV and SVB, since they view the problem from a more broad perspective. Second, UWV and SVB are at the core of the issue. They are key players and have the ability to change internal ways of working, including policy surrounding AS as well as the use of algorithmic systems themselves.

Third, there is a group of supervisory authorities and the client council, who can be seen as either key players or subjects. They have some power that stems from their legal rights as supervisors, and are interested in cases that relate to their supervisory task. The client council is involved through the SUWI-law and has a legal connection to SVB, UWV, and social affairs, and can therefore be seen as a key player. Court of Audit and the Dutch DPA pre-emptively research algorithms and data handling, and are therefore more powerful than the ombudsman and human rights institute, who are assumed to mostly investigate after a case is brought to them. The Dutch DPA has specific abilities to do more than just advise, but also impose measures, and is therefore the most powerful of this group.

The DCA has a very higher interest in the case than other supervisory bodies, as it focusses on AS specifically. However, the DCA has fewer legal instruments to influence the case, and is therefore a subject in the case. The fourth and last group are the NGO's. The two NGOs are both also subjects. Of the two, Amnesty seems to be more interested in the problem, based on the recent campaigns and publications. Due to their size and international nature, they are also slightly more powerful than bits of freedom.

4.6 Confront initial problem formulation with the findings

4.6.1 General conclusions: cooperations and formal powers

Several conclusions can be drawn with regards to the problem at hand. The Actor Analysis provides insight in to the basic playing field within which AS are used and resulting harms are managed. Different supervisory agencies have an eye on the issue within their respective roles, but DCA, as a supervisor on AS specifically, currently holds no formal powers. This can impact the way in which solutions are found. As Enserink et al. (2022) mentions, the different viewpoints and interests in the actor field means solutions to problems need to be negotiated or imposed by a an actor with the power to do so. In this case, DCA does not have power to impose solutions. Remaining actors thus need to negotiate solutions, which can be complicated by plurality of objectives within these actors (see also 4.6.2).

The analysis of interdependencies highlights groups of actors that might work together within the problem situation. However, there are already several cooperations in the field that fill the potential for possible cooperations. The EU, Ministries and Chamber of Representatives hold formal relations that can be seen as a form of cooperation, mainly through law making and accountability. Executing agencies work together in the Manifest group. UWV, SVB and SZW have formal relations as part of the SUWI-legislation. Lastly, the different supervisory agencies work together in cooperations mentioned under the mapping of formal relations (4.3).

4.6.2 Double binds

The 'Tragic double bind' that was mentioned in the introduction can be extended to more double binds, that can also be seen as ambivalence within parties. An important one is the ambivalence within parties in the house of representatives, some focus more on the detection and measures against abuse of the SUWI-system, others focus more on the human dimension and the possible impacts of AS on individuals and society. This ambivalence is also present within ministries and executing agencies. Figure 6 displays 'double binds', including the double bind between governing algorithms and governing by algorithms (Kuziemski & Misuraca, 2020). Others are carefulness versus efficiency, serving citizens versus fighting fraud, and fitting solutions versus equality (Tijdelijke Commissie Uitvoeringsorganisaties, 2021).

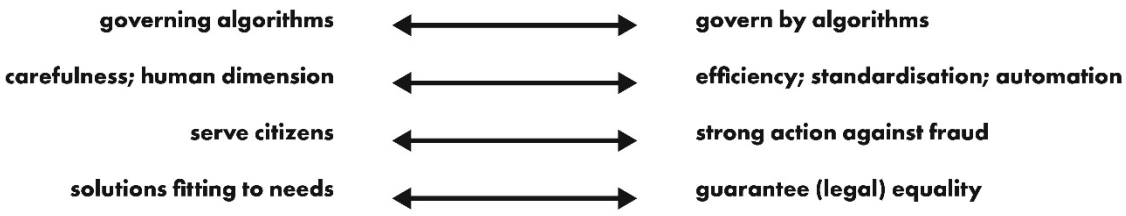

Figure 6 Double binds of actors

Trade-offs made between the two sides of the double bind can shift over time. In recent years, political goals led to focus on efficiency, for which digitalisation was a tool. This was followed by decreasing abilities to find solutions fitting to needs and an increasingly strong focus on action against fraud, partly driven by publicized fraud cases, eventually leading to scandals that tipped the scale towards the human dimension (Parlementaire Enquêtecommissie Fraudebeleid en Dienstverlening, 2024; van Atteveldt et al., 2024). This attention can resulted in publications by actors in the actor analysis, and the start of DCA. It remains to be seen, however, if the intentions and publications of actors, especially SZW, UWV, and SVB, coincides

with a change of the culture that led to scandals in the first place. Attention for the human dimension after the childcare benefits scandal might not yet have led to softening of a harsh repressive climate (van Atteveldt et al., 2024).

4.6.3 Knowledge gaps

The Actor Analysis was unable to clarify on certain relations, viewpoints and actions that thus remain a knowledge gap. The main knowledge gap is the exact oversight over UWV and SVB as documented in the SUWI-law. This oversight is divided between SZW and the labour inspection, but the analysis did not uncover how this is formally organised, especially on the topic of AS. This also led to SZW being considered a non-dedicated actor, whereas they might, in reality, be dedicated more than public sources would show. The same applies to for the National Client Council and Ombudsman.

5. Systems Safety Analysis

This chapter describes the System-Theoretical Process Analysis (STPA) that is done on the governance of AS at UWV and SVB in order to answer the following research question:

How can a Systems Safety Analysis be used to describe algorithmic systems at agencies executing the SUWI-law, and the potential hazards of these systems?

STPA is form of Systems Safety Analysis. What follows is an introduction to relevant systems safety concepts, followed by STPA.

5.1 Systems Safety Fundamentals

In this report, STPA will be done, which is based on Systems-Theoretical Accident Model and Processes (STAMP), developed by Leveson (2012). As described in the introduction, accidents are undesired loss events, where losses are undesirable effects. Hazards are system states that, together with a set of worst-case environmental conditions will lead to accidents and thus losses. Under STAMP, safety is then described as the freedom from unacceptable losses, and is an emergent property stemming from interacting system components. These components can be controlled by imposing constraints on both the components and interactions between them. In this way, safety is framed as a control problem, where control is meant to impose safety constraints.

Different components of a system are displayed in a hierarchical safety control structure (SCS). Each level imposes constraints on a level beneath it. These constraints are imposed through controls, which can be physical design, processes, or social controls. The control actions are called the reference channel. From lower levels there is also a measuring channel back to higher levels, which is feedback about effectiveness of controls (and thus constraints). This general structure is displayed in Figure 7.

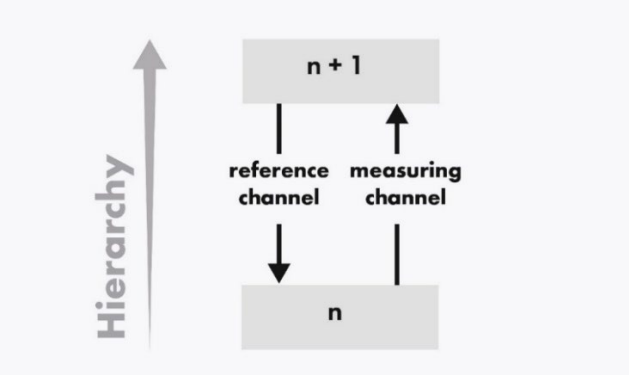

Figure 7 Example of relations in Safety Control Structure (SCS)

An important part of an SCS are the processes controlled by a controller, between which and the process there are again control actions and feedback. A controller also has a process model. For human controllers they are mental models, for automated controllers they are control logic. For a controller in a SCS that is controlling a process, four conditions are required;

1. Goal condition: the controller must have a goal, the safety constraints a controller wants to enforce;
2. Action condition: the controller must be able to affect the state of the controlled process, done through control actions;
3. Model condition: the controller must be or contain a model of the controlled process (the process model);
4. Observability condition: the controller must be able to ascertain the state of the controlled process, the feedback or measuring channel.

These different conditions and the related structure of an SCS are displayed in Figure 8.

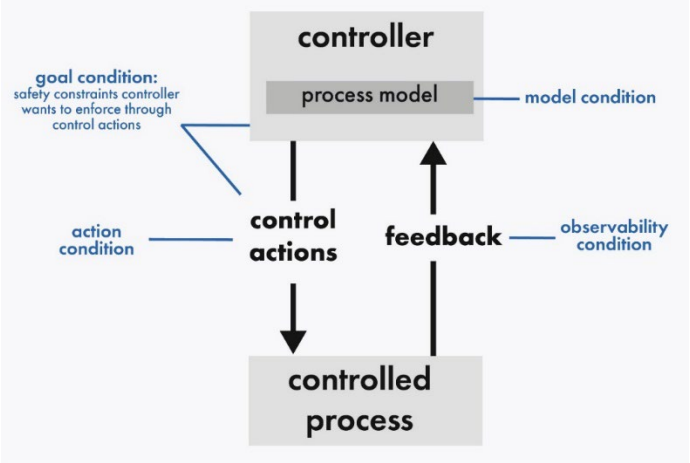

Figure 8 Control conditions in the Safety Control Structure

If the four conditions are not fulfilled, control can become inadequate, leading to accidents. Often these accidents happen when the process models used by the controller are inadequate, leading to inadequate control actions. This can, for instance, happen when feedback is missing. Control can be inadequate in four ways;

1. Control actions are incorrect or unsafe (providing causes hazard);
2. Required control actions are not provided;
3. Potentially correct control actions are provided at wrong time (too early/too late);
4. Control is stopped too soon or applied too long.

A SCS displays the different components in the safety system, which are kept in balance through feedback control loops, creating safety. In order to ensure systems are safe, a SCS must be designed wherein this balance is maintained and safety constraints are effectively enforced.

5.2 System-Theoretical Process Analysis (STPA)

System-theoretical Process Analysis (STPA) is a hazard analysis method based on STAMP used to investigate accidents before they occur. This method includes more causal factors that contribute to accidents, namely social, organizational and management factors. To this end, humans in STPA can be treated in same way as automated components. However, human controllers have dynamic control algorithms that are adjusted due to feedback and changes in goals, versus static control algorithms of automated controllers. Furthermore, a human needs a process model not only of the controlled process, but also of the automation used. If, for instance, an employee (controller) uses an AS (automation) to award benefits (process), the employee must have a model of how benefits are awarded, but also of how the AS works and affects the process of awarding benefits.

STPA, following the STPA Handbook by Leveson and Thomas (2018), consists of four steps. These steps are displayed in Figure 9 and will be followed in the following paragraphs. The different steps of the analysis used information elicited during the Actor Analysis and interviews. For each step, this connection is described.

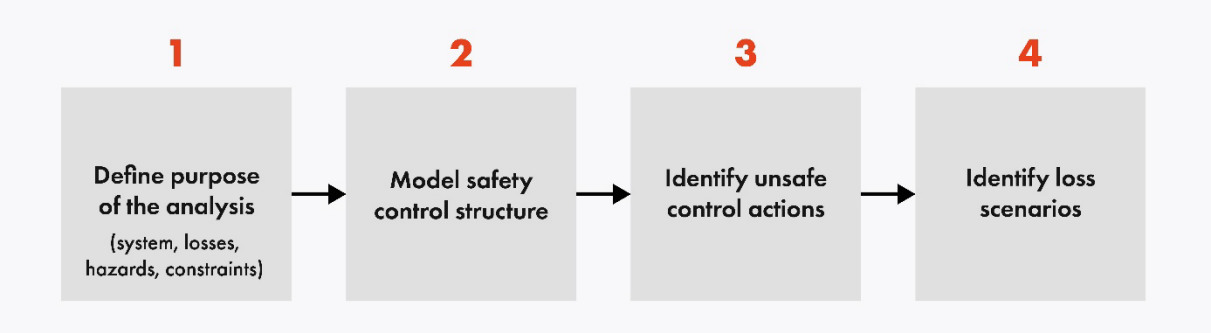

Figure 9 Steps of a System-Theoretical Process Analysis (STPA)

5.3 Purpose of the analysis

5.3.1. System of interest

The system of interest is described under the Actor Analysis. It encompasses processes at UWV and SVB that lead to citizens getting benefits and the oversight over the use of these benefits, for which AS are used. In this analysis, three lenses will be used. The first two look within UWV and SVB, and the third is the larger landscape within which these two parties reside, referred to as the 'general' overview. This general overview is a new addition to STPA, meant to bridge the scope difference between STPA and Actor Analysis (see paragraph 3.2).

5.3.2 Losses

Under STPA, losses result from accidents, which is 'an undesired or unplanned event that results in a loss' (Leveson, 2012). What exactly can be considered a loss can vary, but losses are unacceptable to stakeholders. Following the STPA Handbook by Leveson and Thomas (2018), losses were elicited using the involved stakeholders: by looking at their *stake*, what they find unacceptable losses can be found. In this analysis, the focus is on losses for (groups of) citizens. Losses for organisations, such as financial and reputational damages, have been left out of scope. This resulted in the following losses:

L1 Citizen unjustly uses benefits (malicious)
L2 Citizen unjustly uses benefits (accidental)
L3 Citizen does not use benefits they are entitled to
L4 Incorrect appreciation of benefits
- L4.1 Citizen is unjustly denied benefits (perceived non-entitlement)
- L4.2 Citizen receives too little benefits

L5 Unjust or disproportionate investigations and measures
- L5.1 Citizen is unjustly subjected to investigations
- L5.2 Citizen is disproportionally burdened by investigations (e.g. high frequency)
- L5.3 Citizen is unjustly excluded from benefits (measure)
- L5.4 Citizen is unjustly fined (measure)

L6 Citizen is not efficiently helped (time and effort of citizen)
L7 Citizen is unable to exercise their rights

5.3.3 Hazards

Hazards are described as 'a system state or set of conditions that, together with a particular set of worst-case environmental conditions, will lead to an accident (loss)' (Leveson, 2012). Hazards are used to describe which system states should be avoided in order to prevent losses from occurring. This analysis focusses on three hazard categories, henceforth 'hazards', that are a conceptualisation of problems concerning AS that have been identified in the Actor Analysis and the interviews. The hazards apply to the different types of AS in focus. In all hazards, 'flawed' relates to either incorrect, incomplete, or inaccurate. The following hazards are considered:

H1 Flawed (incorrect, incomplete, inaccurate) logic is used in algorithmic systems
Related losses: L4, L5, L6

H2 Flawed (incorrect, incomplete, inaccurate) information is used in algorithmic systems
Related losses: L1, L2, L4, L5

H3 Citizen has flawed (incorrect, incomplete, inaccurate)understanding of benefits system and role of AS therein
Related losses: L2, L3, L7

The first type of hazard relates to model that is used of reality, i.e. the logic used in the different AS. The AS are applied to make decisions about real life situations, or selections in an actual group of people. They therefore hold a certain logic that is used to make these selections and decisions, built on assumptions to create approximations of reality. These assumptions can be based on laws, regulations, (historical) data, but also viewpoints of those making the models or society at large. If the assumptions are a flawed approximation of real life complexity, the models built on these assumptions can create losses. This hazard relates to reporting about AS being unable to account for complex cases (e.g. Tijdelijke Commissie Uitvoeringsorganisaties (2021), Peeters and Widlak (2023)), but also recent cases wherein risk assessments unjustly or disproportionately flagged certain groups (e.g. DUO, Childcare Benefits Scandal). Furthermore, interviewees pointed to mismatches between logic used in AS and reality, leading to issues (see I1, I5, I6, I7), and the Actor Analysis pointed towards complexity as an important issue (see Appendix A.2).

The second hazard relates to the model used of a citizen, i.e. the information that is used about a citizen. Information about a citizen is used to judge whether they are entitled to benefits, how much they should receive, but also to judge whether they possibly use benefits they are not entitled to. If information is flawed, this can lead to losses. Citizens intentionally providing flawed information also falls under this hazard. Use of flawed logic (H1) can lead to exclusion

of important information, and thus cause this hazard to occur. This hazard relates to discussions on data exchanges within government and possible consequences of flaws therein (e.g. Stichting Kafkabrigade (2023) or van Atteveldt et al. (2024)), also pointed towards by interviewees (see I1, I7) and in the Actor Analysis (see Appendix A.2). It also ties to abuse of systems by citizens to provide flawed information on purpose (see I3, I7 and Appendix A.2).

The third hazard relates to the image a citizen has of the system in focus, that is the processes executing agencies use to award benefits, enforce rules, including the use of AS therein and the influences citizens have for going against decisions or altering decision-making. This was also pointed out by interviewees (see I1, I3, I6, I7), and was found to be an important theme amongst actors in the Actor Analysis (see Paragraph 4.2 and Appendix A.2). If a citizen does not properly understand what the benefits system looks like, this can lead to losses.

5.3.4 Constraints

In Table 4, the list of possible hazards is displayed, together with related system-level constraints needed in order to prevent the hazard. System level constraints are what needs to happen to prevent hazards and thereby also resulting losses. System level constraints listed in Table 4 are overarching constraints that are assumed to be generally accepted. They have been elicited from European and Dutch policy documents (European Commission Directorate-General for Communications Networks Content and Technology, 2019; Rijksoverheid 2022; van Huffelen, 2022a).

Table 4 Hazards and overarching system-level constraints

Hazard	Related system constraints
1 Flawed logic is used in AS	• Prevention of arbitrariness • Prevention of discrimination (direct/indirect) • Limitation of bias (data, logic, humans) • Exclusion of citizens must be prevented
2 Flawed information is used in AS	• Data quality must be upheld • There must be ownership of data
3 Citizen has flawed understanding of benefits system and role of AS therein	• There must be active transparency • There must be passive transparency • Information is understandable for citizens

Constraints for the first hazard include preventing arbitrariness, discrimination, bias and exclusion of citizens. Discrimination regards both direct discrimination and indirect discrimination, via proxy variables. Discrimination can be understood as an outcome of use of AS with biases present in them in some way (Balayn & Gürses, 2021). Bias can be present in the data that is used (skewed data), in the logic in the AS (connections between and weight of variables), but also in human controllers, both those developing the AS and those using it. All these forms of bias relate to each other. While it is clear bias and discrimination must be prevented, much can be said about what the terms mean, and what is to be done to prevent them (see e.g. Balayn and Gürses (2021); European Union Agency for Fundamental Rights (FRA) (2022) and Schwartz et al. (2022)).

For the second hazard, the main constraint is that of maintaining data quality. Data quality is a term that can include various factors. Most often, the understanding of data quality includes accuracy (syntactic and semantic), completeness, consistency, and timeliness (Batini et al., 2009). In addition, it is important that those using the data (i.e. executing organisation) are able to ascertain ownership over data in order to be able to be accountable for decisions based on it.

For the third hazard, it is important that there is transparency to both those who may be affected by AS and those who must be able to oversee their use. This can be split into active transparency, for example motivations given to citizens on decisions or reporting to politicians or supervisors, and passive transparency, for example information that is findable on websites or registers when looked for. An additional constraint would be that this information is understandable to citizens. Similar to previously mentioned constraints, these constraints are general and much has been written about what they mean in theory and practice (see for instance Diakopoulos (2020) and Wieringa (2020)).

5.4 Safety Control Structures

Three Safety Control Structures (SCS) were constructed, one each for UWV and SVB, and one that shows the overarching supervisory landscape surrounding these organisations. The intraorganizational structures were made using the algorithm registers these organisations have published, identical to information published on the national algorithm register. The selection of AS taken into account can be found under Table B 1 (Appendix B). This was then expanded with publicised reports and documents. The general SCS was built with the formal chart (Figure 4) as a basis, replacing formal relations with corresponding control actions and feedback, and adding or replacing actors with components.

The individual SCSes of UWV and SV are displayed in Figure 12, Figure 13 and Figure 14 . Several visual additions have been made in order to make these diagrams more easily readable. A legend for these elements can be found in Figure 10. Processes that aim to awarding benefits are coloured green, those that aim to enforce rules are orange. Components within the organisations are in white boxes, while external components are in blue boxes. Lastly, due to the amount of components and interactions between them, several lines have been cut to increase visibility. Those have been connected to grey boxes to mark the origin of the connection.

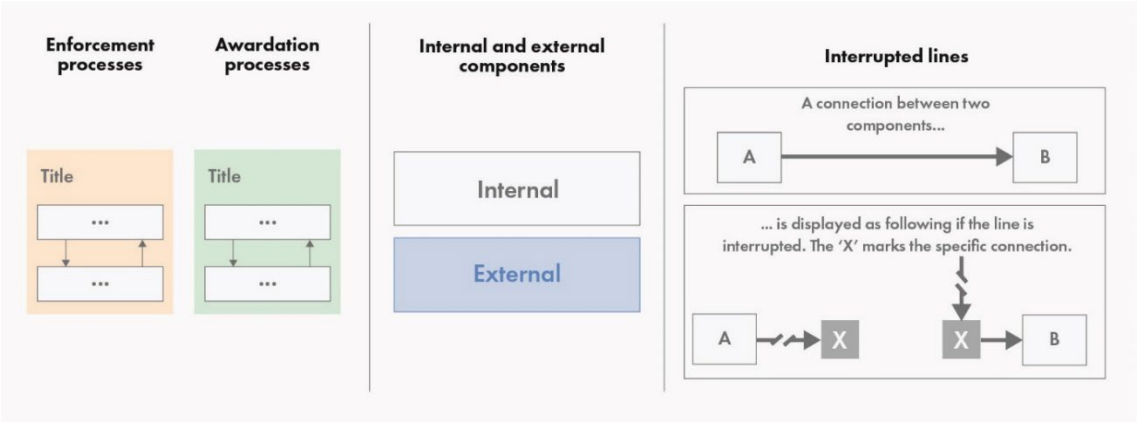

Figure 10 Legend for SCSes of UWV and SVB

5.4.1 General safety control structure

The general SCS, shown in Figure 11, displays relevant components relating to use and supervision over AS outside the boundaries of SVB and UWV. This SCS is a new artefact as it shows inter-organisational relations, while the scope of STPA is typically intra-organisational.

Supervision of execution of SUWI-laws lays primarily with the executing organisations, who report to SZW about their activities (Ministerie van Sociale Zaken en Werkgelegenheid, 2021). Additional oversight by SZW is risk-based, using signals from Ombudsman, client councils, unions, employee councils, and reporting by UWV and SVB themselves. This includes yearly reports and 'bottleneck letters', providing overviews of bottlenecks found during execution. SZW provides budgeting, law, and policy that influences UWV and SVB, sometimes directly leading to development and use of AS (Uitvoeringsinstituut Werknemersverzekeringen, 2021b, 2022). Law and policy is influenced by feedback from UWV and SVB, but also by the Chamber of Representatives. This Chamber receives the justification reports from executing organisations through SZW, can ask formal questions, make, change, and approve of policy. BZK is responsible for digitalisation policy that influences UWV and SVB. They started the algorithm register, and appointed AP to house the national algorithm coordinator, resulting in DCA. DCA currently has no formal ways to exert control over other organisations or receive feedback from them. However, there is assumed to be discussion between parties, and DCAs risk reports might impact organisations. As part of their digitalisation policy, BZK made an implementation framework, and is working on a broader and more cooperative framework (Digitale Overheid, 2023; Ministerie van Binnenlandse Zaken en Koninkrijksrelaties, 2023b). CRM and ARK have also published frameworks (Algemene Rekenkamer, n.d.-b; College voor de Rechten van de Mens, 2021). Ombudsman has no framework, but released viewpoints on AS-use by government (Govers et al. (2021)), which can similarly impact organisations.

Citizens can directly complain to CRM when it comes to discrimination, or to Ombudsman regarding escalated complaints. CRM can investigate and make non-binding decisions, Ombudsman can provide non-binding advice. Both report on topics relevant to their work, and advise on laws and policies. These organisations are also part of two feedback loops, where policy created higher up influences citizens through actions by executing agencies, and these influences come to Ombudsman and CRM through complaints and investigations, and are then fed back to politicians through reports, judgments and advices. The Court of Audit investigates organisations state-wide, and published general reports yearly, and special reports intermittently. It is part of a similar feedback loop, although this loop is smaller: the Court of Audit investigates organisations, not complaints from citizens. Feedback from citizens to higher up levels in the SCS is also processed through NGOs that lobby and appeal to political actors. Furthermore, the National Client Council represents citizens and passes along feedback to SZW, for instance through bottleneck letters, similar to SZW and UWV. The press has an important role in the SCS. They can investigate what happens at executing agencies and have in recent years reported on scandals at the executing organisations. Information requests filed by the press led to publication of internal documents that shed light on governance of AS. In the general SCS, the output of the press is linked to a cloud symbol, as is the output of the algorithm register. This is done because these outputs are not control actions or feedback specifically fed to one or several other components. In short, the outputs are related to whomever reads them, which can be individuals present within all mentioned organisations or groups in the SCS.

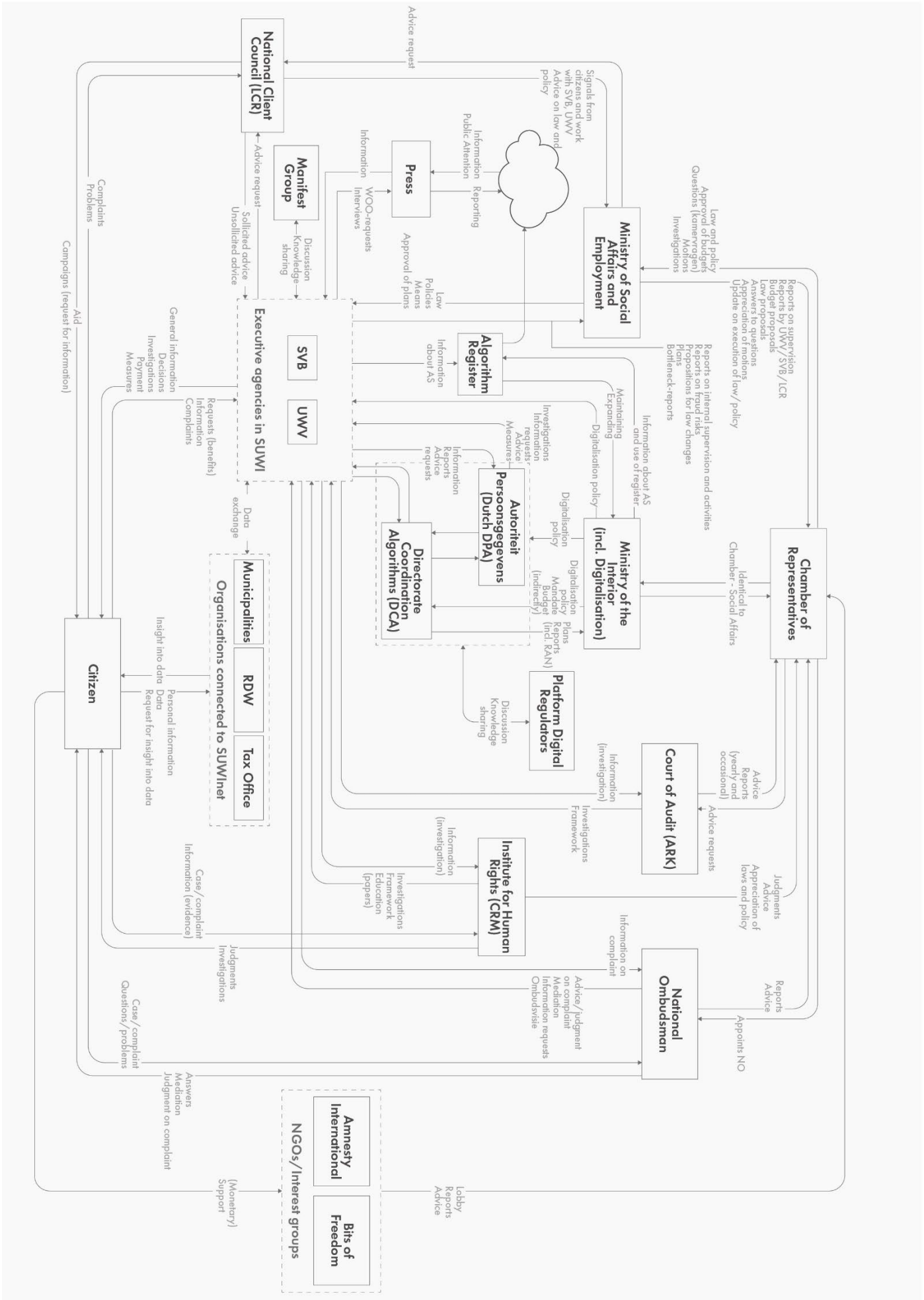

Figure 11 General Safety Control Structure

5.4.2 UWV safety control structure

The safety control structure depicting UWV can be found in Figure 12 and Figure 13. Two risk scans were included, one of which has been discontinued, but was still part of the register at the time this research started. Risk scans provide seventy percent of signals that get sent to an employee for investigation, the remaining thirty percent being random so the employee does not know whether a signal resulted from a risk scan or not. Since 2021, UWV uses the Model Risk Management policy (Uitvoeringsinstituut Werknemersverzekeringen, 2021a). This policy provides a three lines of defence structure with formalised roles and interactions. A development team within the enforcement division is responsible for development and maintenance, and has the ability to seek advice from an ethical committee consisting of internal and external members, the UWV client council, and an external validator. The policy describes this is to be done during development. The model owner of a risk scan is part of the division it is used within, in this case the benefits payment division. Frequent monitoring is done by a business control and quality team within this division. A workflow system and dashboard holding outcomes of investigations are used for monitoring, bias control and optimisation (Uitvoeringsinstituut Werknemersverzekeringen, 2021b). Requests for unemployment benefits are automatically sent through the risk scan culpable unemployment. During development of this risk scan, a data analyst sent cases from the risk scan to executing employees through a digital environment (Uitvoeringsinstituut Werknemersverzekeringen, 2022). After development, this process was supposed to be further merged into the process of awarding benefits, although no documentation was found to depict this new situation. It was therefore depicted with a data analyst in the loop. Even in this form, it shows the coupling of awarding and enforcement processes.

The other AS included in this report are used for awarding benefits rather than enforcement. Fewer information was found on these AS. The algorithm register mentions specialised personnel that helps with monitoring and changes, and this is assumed to be structured similarly to risk scans: a business control and quality team monitors the AS and use thereof, and there are developers that develop and maintain the systems. The client application unemployment benefits can directly award benefits and takes information from external sources, assumed to be through SUWInet. Difficult cases, together with some random cases are passed to an executing employee. The Claim Beoordelings- en Borgingssysteem System (CBBS-system) serves as a guide for a trained executing employee to establishes a disability percentage that influences the height of benefits. This is done by judging to what extent a citizen is able to work, for which the CBBS-system holds a list of job descriptions. Outcomes of judgments are monitored at random. The digital sickness checklist is used to judge whether someone should receive sickness benefits. The AS provides a personalised questionnaire for someone applying, which is then used by an executing employee to judge the application. An external organisation checks for ethical and legal norms on top of regular outcome checks.

5.4.3 SVB safety control structure

The safety control structure of SVB can be found in Figure 14. Five AS have been included, two of which are used for enforcement. One is a decision tool that helps an employee decide whether a measure should be taken and how high this should be. Another is the SWAN-model, which is not in use yet but already mentioned in the algorithm register. It is a self-learning risk scan that aims to signal possible misuse of AOW. It learns from cases where abuse was found.

Random signals will be added to the signals sent for investigation, similar to UWVs 70/30-rule (Hamer et al., 2022).

Three other AS are used to award benefits. The systems for elderly pension (AOW) and child benefits are separate AS, but have been modelled together due to their similarity. These are designed to automatically process a majority of requests. The remaining difficult cases are judged by employees. These AS make use of external databases, again assumed to be through SUWInet. Similar to the digital sickness checklist of UWV, the AS used for child benefits provides a personalised request form that is then filled in by the citizen. The AS used for additional pension can make decisions for a citizen, but requires intermediary decisions by an employee.

Fewer is known about SVB, as fewer information requests were found. Development is assumed to lay within the IT department, which can ask for advice from citizen panels if impacts on citizens are expected (Hamer et al., 2022). SVB has an ethics centre, and uses Intention Workshops (Garages de Bedoeling), where employees can discuss ethical problems they encounter in the execution of their work. Based on these workshops, changes to internal policy and external policy are proposed, practical solutions are found, and tools are made (Berg et al., 2019).

5.4.4 Similarities and differences between UWV and SVB

The two safety control structures of UWV and SVB can be compared, resulting in several differences and similarities. Both executing agencies make use of random samples in their risk scans, in addition to flagged signals. This is ongoing practice for UWV, and will also be included in the risk scan SVB has in development (Hamer et al., 2022). Both agencies have organised governance surrounding privacy risks, although this is mandatory for governmental organisations under the GDPR. On the area of ethics, both organisations have policy and advisors, although organised differently. As part of the same law, the structure of policy coming down from SZW and reporting going back is the same.

There are also notable differences in the approach. The ethics centre of SVB is internal, whereas the ethical committee of UWV is a combination of both internal and external members. Their functioning within the organisation is also different. The ethical committee can give (un)solicited advice on AS and surrounding policy. The ethics centre, however, takes on a more active role in SVB, through organising workshops, proposing internal policy, and developing tools for employees. Comparing this way of working, where policy is adjusted through bottom-up workshops, to the formalised model risk management policy and ethical compass of UWV, it could be seen as a signal of two different approaches. UWV taking a top-down approach, SVB a bottom-up strategy.

A last notable difference is the placement of the enforcement side of the organisation. SVB has placed their enforcement department under the service division, at UWV the enforcement side is a separate division. However, as UWV includes enforcement of rules as part of awarding processes, as can be seen with the culpable unemployment scan, both organisations can be said to entangle enforcement and awarding to some extent, although in different ways.

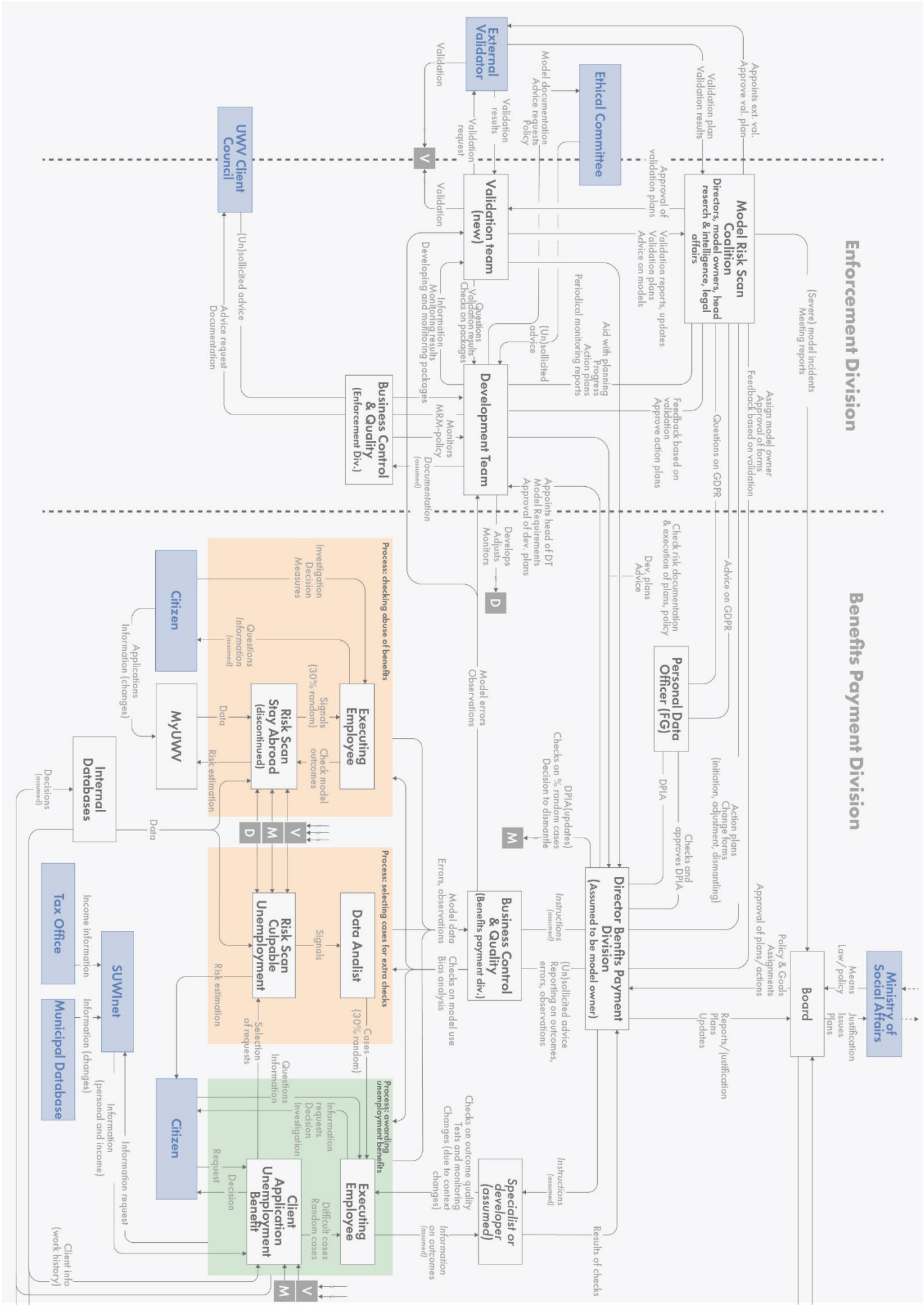

Figure 12 Safety Control Structure UWV

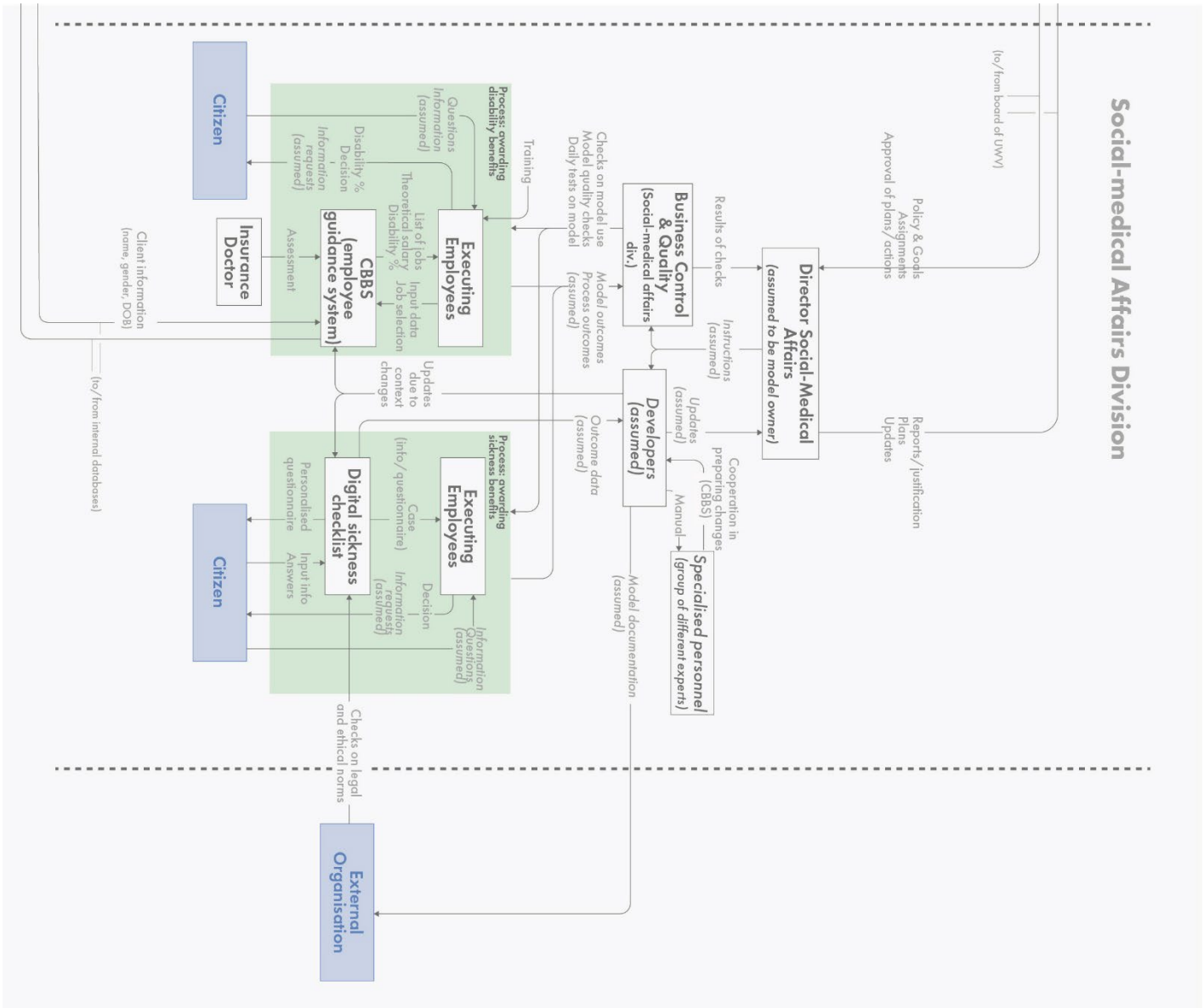

Figure 13 Safety Control Structure UWV (continued)

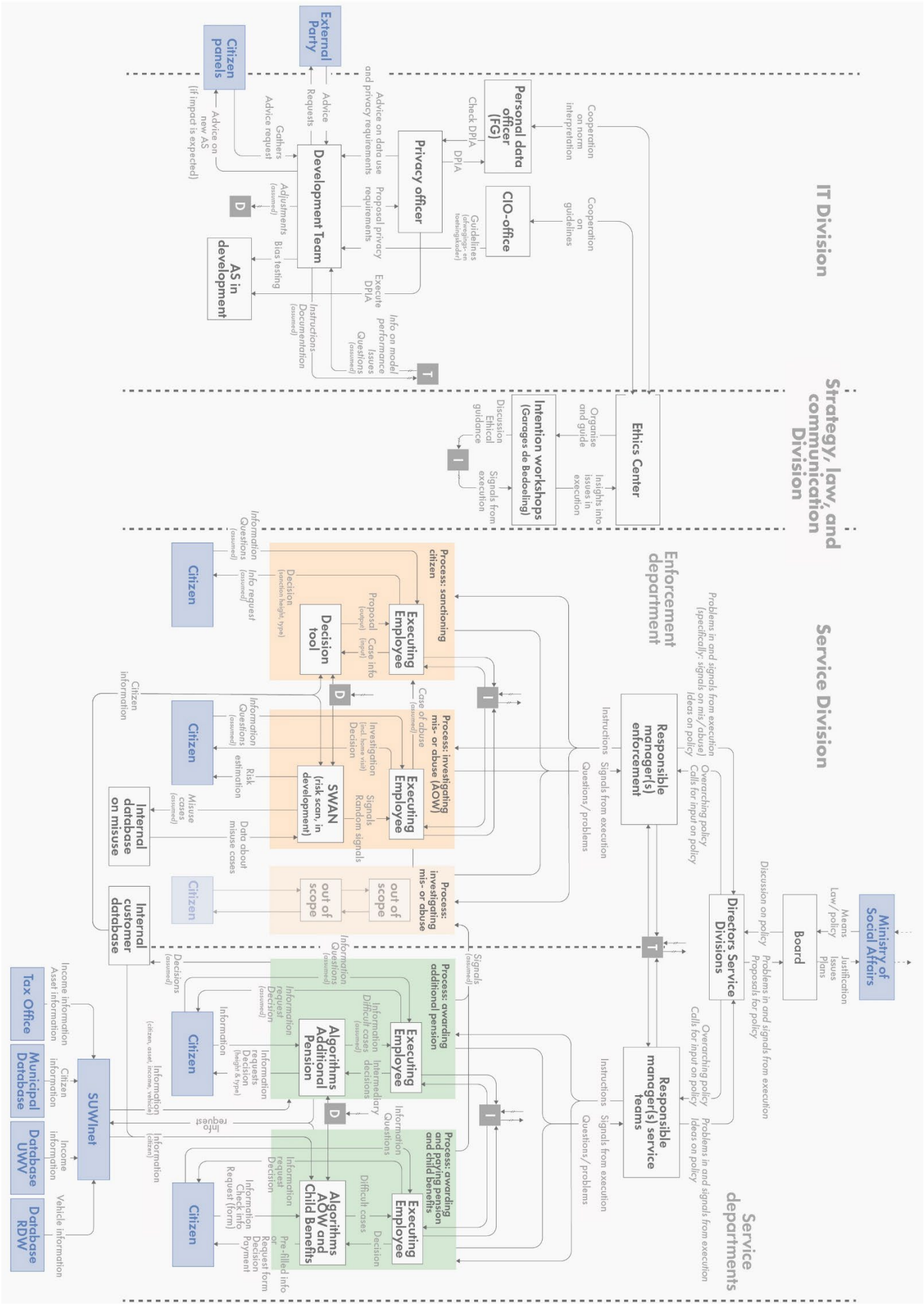

Figure 14 Safety Control Structure SVB

5.5 Identify potentially hazardous control actions

The documented safety control structures can be used to explain how hazards can occur. Control actions can be inadequate in the four ways described under paragraph 5.1. For each of the three hazards, the different control actions that can contribute to the hazards have been identified. This can be found under appendix B.2. What follows is a description of important control actions that can lead to hazards. The control actions have been divided into two categories. The first are control actions that outside the controlled process, that can be considered to relate to systemic or indirect factors in the coming to be of hazards. The second category is control actions in the controlled processes. These can be considered direct factors.

5.5.1 Control actions surrounding processes (systemic and indirect factors)

Law and policy creation by chamber of representatives and ministries (H1, H2, H3)

In the same way as internal policy influences control actions within organisations, law and policy created higher up in the SCS influences internal policy and supervision of the organisations. Laws and policy can cause hazards if they are insufficient or not well explained. They can also cause hazards due to stacking, which increases complexity and thereby also understandability. Not providing laws and policy also creates hazards, as it does not enforce wat is needed, or creates gaps in supervision.

Reporting by the press (H1,H3)

Reporting by the press can be an important extra set of eyes on AS use and possible issues. It can uncover flawed practices at organisations, describe impacts on citizens, and uncover scandals. Not providing it thus causes hazards. However, providing it can also cause hazards. It can draw focus on scandals and role of AS, obscuring other relevant factors. Furthermore, a focus on abuse of social benefit schemes can cause strong reactions that in turn can lead to new scandals. Reporting can also instil fear with citizens, who might be afraid for repercussions if they make mistakes. Reporting thus influences politics, citizens, and executing agencies

Requesting advice from external parties: citizen representation, ethical committees, external organisations (H1, H3)

Asking for advice from external parties can cause hazards when the requests are scoped in such a way that insight from them is limited. For instance, when the focus is on ethical or legal aspects more than impacts on citizens. Not requesting advice can also cause hazards, as it limits insights into citizens' reality, understandability and other factors. External advice is often seen in context of development of AS, whereas insights from external parties can create valuable insights in later stages of the AS lifecycle. In that case, this control action is stopped too soon. Inadequacy of this control action can lead to missing or limited feedback from external parties.

Monitoring of model outputs and use (H1)

Monitoring the model itself and usage thereof is important to spot (potential) errors and oversee model performance. It provides important feedback for management and development and maintenance. Not providing this control action thus leads to hazards. However, providing this control action with a focus on KPIs or technocentric monitoring can obscure important factors that are not taken into account. This can cause flawed logic to persist and feedback to be limited.

Development and maintenance of AS (H1)

Development and maintenance of AS is an important control action to ensure proper logic and adapt AS to improve logic or better fit to context. Not providing maintenance can cause hazard, as it is needed for improvements and to reflect context changes. However, if there is limited insight into relevant factors present with developers, this can cause flaws to be translated into the AS. These can persist during maintenance if mental models are not properly adjusted (e.g. through external advice, instructions, or monitoring). If maintenance on systems is done too late, this can also cause hazards. Asynchronous development of laws, values and technology can cause AS to not properly represent current values and laws.

Information provided through register (H3)

The information that is not actively given, but can be found through the algorithm register influences ability of citizens to understand processes and involvement of AS therein. Not providing information through the register causes citizens to have even fewer information. Giving information too late can also cause hazards, as it is possibly no longer actionable and citizens might already be in dire situations, such as debt, in which it is harder for them to act.

Automated data transfers in SUWInet (H2, H3)

Automatically transferring data through SUWInet can increase complexity, making it less understandable for citizens. It can also cause permeation of flawed information through different systems. Information can have different meaning in different systems, or be less current than is needed. This can cause effects crossing boundaries of organisations.

Instructions from management and internal policy (H1, H2, H3)

Instructions and internal policy influence all aforementioned control actions. It can influence the development, use, and monitoring of AS. It can influence if and how external parties are consulted and what information is given to citizens. It can also limited or expand discretionary space for employees, and cause or prevent overreliance on AS. Providing instructions and policy can thus cause hazards. Not providing them, on the other hand, can also cause hazards, for instance due to a lack of scrutiny or arbitrariness.

5.5.2 Control actions in controlled processes

The different SCSes under paragraph 5.4 display several processes. In Figure 15, general archetypes of these processes involving AS are displayed. These different processes under the analysis in this research cause the general hazards to translate into losses. The different processes show different relations between employees and AS, but also between AS and citizens. This is of importance to how hazards translate into losses. The following control actions are part of these processes, and can be considered direct factors.

Decisions (H1, H2) and Information provided to citizens (H3)

Decisions can influence eligibility and height of benefits, but also the type of measure that is chosen (SVB uses a decision tool for sanctions, see Figure 13). Decisions are either made directly by AS (SVB uses ADM) or by an executing employee using AS, when this is company policy, for difficult cases, or when it is a decision tool or risk scan. The AS then provides a employee with a case or advice. Decisions based on flawed information or flawed logic used in the different types of AS can cause harm to citizens, either directly (ADM) or through executing employees. Decisions can involve an incorrect appreciation of benefits, and thus can

ensure the hazard becomes a loss. Making decisions late causes citizens to be inefficiently helped. Decisions will also involve information on the decision made to be provided to citizens, either automatically generated or provided by an executing employee. This information can impact ability of citizens to understand why decision was made and how AS was involved, or where information came from. Providing information late can render it useless in certain situations.

Risk estimations (H1, H2)

Risk scans make risk estimations on citizens. In reality, this estimation is done on information about citizens or directly on their requests. These risk scans can be based on flawed understanding and flawed information. Although the signals stemming from risk scans are combined with random cases, they are still transferred to employee and can lead to unjust or disproportionate investigations and measures. Providing this control action can thus lead to harm.

Investigations and measures (H1, H2) and Information provided to citizens (H3)

Investigations and measures can be based on flawed logic or flawed information. This can be due to the risk scan used, which relates to the risk estimations made. It can also relate to the decision tool used. Providing investigations and measures can thus lead to harms to citizens. On the other hand, not providing them, or providing them too late can mean citizens make unjust use of benefits.

Similar to decisions, investigations and measures will involve informing citizens. Two ways, err logic in risk scan (flagged) or in decision tool (harsh). This information can impact ability of citizens to understand why decision was made and how AS was involved, or where information came from. Providing information late can render it useless in certain situations.

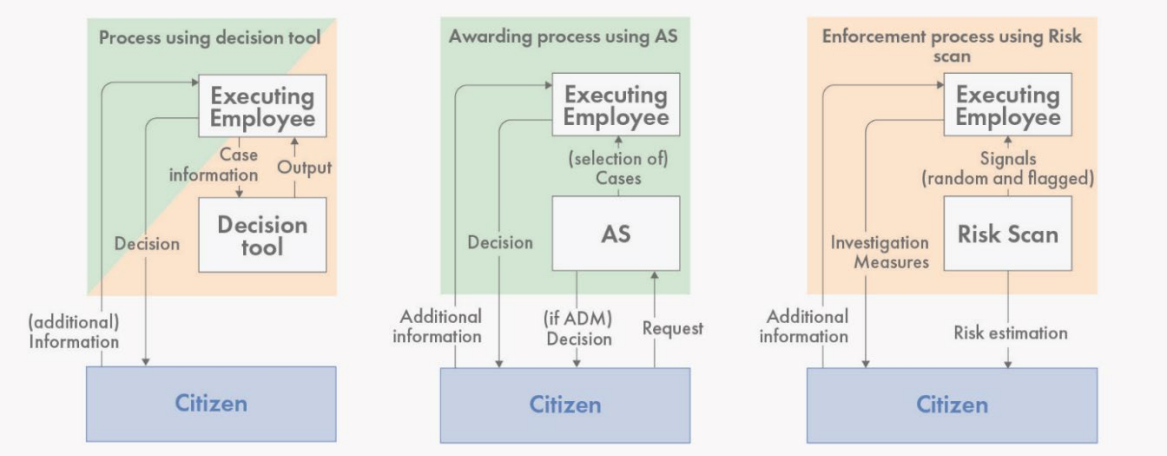

Figure 15 General types of processes involving AS

5.6 Determine how unsafe control actions could occur

Causal scenarios can help build a theoretical understanding of how control actions can lead to hazards, and how these hazards can lead to losses. Four causal scenarios were made, which will be discussed in this paragraph. As displayed in Figure 16, a causal scenario displays several events that lead to hazards and losses. The arrows display the causal relation: event A leads to event B, and so on. Events are related to the identified inadequate control actions.

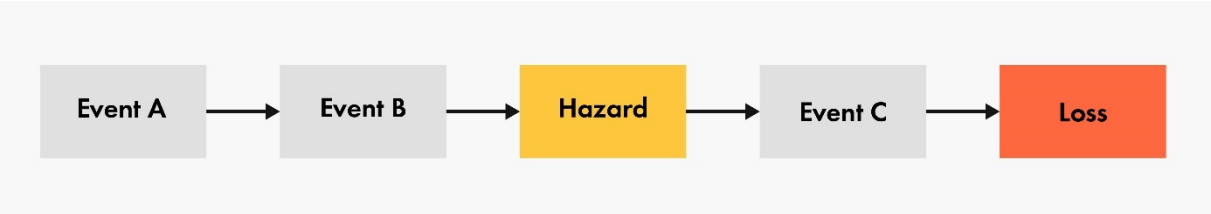

Figure 16 Example of causal scenario

The causal scenarios are initial artefacts that can be used to build a shared understanding of how hazards and losses can occur in sociotechnical systems AS are part of, and what possible actions can be taken to prevent them. They are the culmination of different insights of this research. To construct the scenarios, hazards and related losses were placed together. Next, the identified inadequate control actions were placed around them as events to form causal scenarios with a logical order. The order and relation built on four pillars. First, the feed forward from the previous paragraphs on STPA, in particular the SCSes and knowledge gained while building them, as well as the resulting potentially hazardous control actions. Second, the interviews helped understand how different events tie together and lead to losses. The input of the interviews is briefly discussed for each causal scenario. Further results and a more comprehensive explanation of results of the interviews can be found under Chapter 6. Third, the scenarios built on known relations and cases that were described in research and reporting, as discussed under the descriptions of the scenarios and under the Actor Analysis. And fourth, the causal scenarios are constructed using assumptions and synthesis of these different sources. Construction involved interpretations. As such, they serve as an artefact that can be discussed and built upon to come to a shared understanding.

5.6.1 Causal scenario 1 & 2 : Flawed logic used in AS at UWV and SVB (H1 & L4, L5, L6)

In the causal scenario of UWV, displayed in Figure 17, internal policy leads to citizen representation being scoped narrowly during development, and a monitoring structure that is based on KPIs. This leads to an AS being developed that has a flawed logic, for instance a wrong hypothesis on possible misuse in risk scans or an assumption of reality that is translated into an automated process. Moreover, monitoring with KPIs has limited insight into all important factors that could have been uncovered if representation was spoken to. This is not corrected during use, as consultation of representation when AS are in use is limited. This means feedback to management does not include the right signals and thus internal policy and instructions are not adjusted. The same goes for feedback to developers, which means the flawed logic is not adjusted. The hazard of flawed logic means risk scans can unjustly or disproportionally target certain citizens (L5), which can in turn lead to unjust exclusion or measures (L5). It also means that standard processes will not match the complex reality of citizens. This either means decisions are made that are not correct (L4), or that decisions need to be made manually, which takes longer (L6).

The causal scenario of SVB, displayed in Figure 18, is similar to that of UWV. Citizen representation is only consulted if impact is expected and only during development, leading to

limited insights in the development and subsequent monitoring. For SVB, however, little evidence was found for a monitoring structure with a focus on KPIs, so this was omitted, although it might be present in reality. It was found, however, that SVB aims to have the risk scan it is developing learn from prior cases of incorrect use of benefits (Sociale Verzekeringsbank, n.d.-c), as is also displayed in the SCS in Figure 14. SVB also has a risk overview, with all known fraud risks (Houtzager & Verbeek, 2022). This leads to a risk of runaway feedback loops, where the outcome of the risk scan is fed back into the risk scan, causing overestimated results (European Union Agency for Fundamental Rights (FRA), 2022). If the risk scan flags citizens with certain characteristics, and these citizens are found to indeed make incorrect use of benefits, then the system will learn from this and confirm the initial characteristics it selected on.

What these causal scenarios display is that if the SCS allows limited insights to be both translated into systems and into the monitoring structure, then monitoring will most likely not uncover the resulting flawed logic. This means flaws and internal policy are not adjusted and the hazard persists. Furthermore, the causal scenario of SVB highlights the risk of runaway feedback loops to develop. Within the organisations, discretionary space and personal contact could help prevent losses, as it can help diverge from the standardised processes and profiles (i.e. the logic) that have been coded into the AS. However, as was described in the introduction, digitalisation can be said to have led to a decrease in discretionary space for employees of executing agencies. Internal policy influences the discretionary space of employees. This policy, in turn, is influenced by laws and policies from political actors. Ministries have, following New Public Management-types of politics, pushed for targets and KPIs (van Atteveldt et al., 2024). The reigning political climate, including NPM, and digitalisation of government can be seen as important systemic factors that contribute to hazards and losses. This can reflect on internal policy, the use of KPIs within organisations, and further limit the discretionary space employees have.

Solution directions

The causal scenarios allow for several solution directions to be drawn up. Moreover, the SCS'es that were made allow to place these possible solutions within the structures. A first and apparent solution direction is increasing the involvement of citizens in the development, but also the subsequent use of AS. During development, the respective development teams of UWV and SVB already make use of client councils and citizen panels. The choice of representation to involve can vary depending on the type of consultation, although it could be assumed client councils have a more structural knowledge base. What is important, is the framing of the involvement of citizens. UWV sought to involve the client council to discuss understandability of their risk scans, SVB aims to involve panels when impact is expected. These can both be seen as framings that narrow the scope of discussions, and thereby the possibilities to test assumptions and thus logic. A broader framing, both regarding when citizen representation is involved and for what purpose, might help prevent erroneous profiles or lacking standard processes to come into use.

Apart from involvement during development, the causal scenario highlights that the inclusion of feedback from citizens for monitoring purposes can help adapt logic through development teams, but also through changed internal policies. It would make sense to place this within the standard processes, where a feedback loop could be fed back from the citizen. Suggestions, complaints, and questions should be fed back from citizens to executing employees. SVB, with their Intention Workshops, already has this feedback somewhat: executing employees can

bring forward signals, which then influence internal policies and ways of working. A next step is ensuring this feedback is gathered on a more structural basis, not just through observations of executing employees. This does require feedback to be delivered to a team that, altogether, has the right mental model to understand where issues lie and where changes need to be made. That is, knowledge of the workings of the processes, law, and of the workings of the AS.

Increased citizen involvement in development of technology falls within recent developments of technology ethics, and has resulted in practices such as Value-Sensitive Design and Design for Values (Swierstra & Vermaas, 2022). Designing AS asks for involvement of citizens, as law and policy does not account for all factors relevant to them (Delfos et al., [Under Review]). Including citizens, either directly or through democratic structures, can strengthen legitimacy of algorithmic decision-making in government (Grimmelikhuijsen & Meijer, 2022). Moreover, the interviews (see Chapter 6) highlighted further inclusion of citizen representation and societal groups as a possible governance improvement within the SUWI-system.

The issue of monitoring being unable to capture relevant factors lies with the use of KPIs that were made during development, and technical monitoring structures that were set up to monitor these KPIs. Preventing this could be done through decoupling development of the AS and monitoring of the AS to some extent. By providing a monitoring structure that focusses on factors relevant in the process (i.e. enforcement, granting of benefits), a more complete image could be made. This type of monitoring should then be produced (in coordination) with the executing teams. Moreover, it is important that monitoring has the ability to be sensitive to factors outside of set KPIs, and that the factors included can change over time instead of being set once during development. This asks for sensitivity for feedback from citizen (representation) to executing employees, and similar sensitivity for feedback from executing employees to managers and developers.

Runaway feedback loops, such as can exist at SVB in particular, could be prevented by technical solutions, such as editing the sensitivity to feedback. It could also be prevented by ensuring cases that are learned from are properly analysed and understood before they are included. Subsequent (bias) monitoring can help prevent certain groups being disproportionately targeted. It asks for proper understanding of the reality of citizens, for instance to be able to understand certain characteristics show up more often, and to analyse whether this is reasonable or not. However, insights from the effects of runway feedback loops in predictive policing, together with insights on affirmation bias and the effects of enforcement of social security schemes on individuals, might be reason to aim for more caution. This could mean ensuring risk scans never learn from known cases all together.
Lastly, the effects standardised processes that do not align with the reality of citizens can have asks for the ability to diverge form standard processes in due time. This means that citizens must have the ability to get personalised aid in the first place, and preferably before they are ran through a standardised process. Furthermore, it asks for the ability for executing employees to diverge from standardised processes and recommendations by AS. This asks for policy, instructions, and culture that allows for discretionary space, which could ask for decreased focus on KPIs that have pressured executing agencies in the past. It also asks for a good mental model, that is: executing employees must be able to understand both the situation of the citizen, the relevant law and policy, as well as the process (including AS) in order to understand when to diverge from standardisation, and to recognise and understand why a citizen gets stuck and what they are entitled to. Furthermore it asks for limitation of overreliance on AS.

System-level constraints

Initial general system-level constraints were (paragraph 5.3.4);

- Prevention of arbitrariness;
- Prevention of discrimination;
- Limitation of bias;
- Prevention of exclusion.

The possible interventions can be translated into additional constraints that are more specific than these original constraints set under paragraph 5.3.4. These are;

- Citizen representation must be consulted both during development and use of AS;
- There must be feedback channels from citizen (representation) to those monitoring AS, also on other characteristics than those that are monitored on;
- Extent to which an algorithm learns from its own outputs must be limited as to prevent runaway feedback loops;
- For standardised processes, executing staff must have the ability to diverge from these processes.

Input from interviews

Asides from mentioned literature, the Actor Analysis, and the previous paragraphs on the SCS and control actions, these scenarios and possible solution directions build on insights from interviews. In particular, it relates to subthemes 'gaps between technical teams and legal, ethical, political issues', 'good technical monitoring, but difficulty accounting for all relevant factors', and possible improvement '(proactive) connection to citizens and societal groups' (see Appendix E and Chapter 6). A focus on KPIs was also found in I3, although this interview also suggested that there was feedback from complaints outside of KPIs. Interview I7 brought forward mismatches between AS and complexity of citizens, and resulting issues citizens experience.

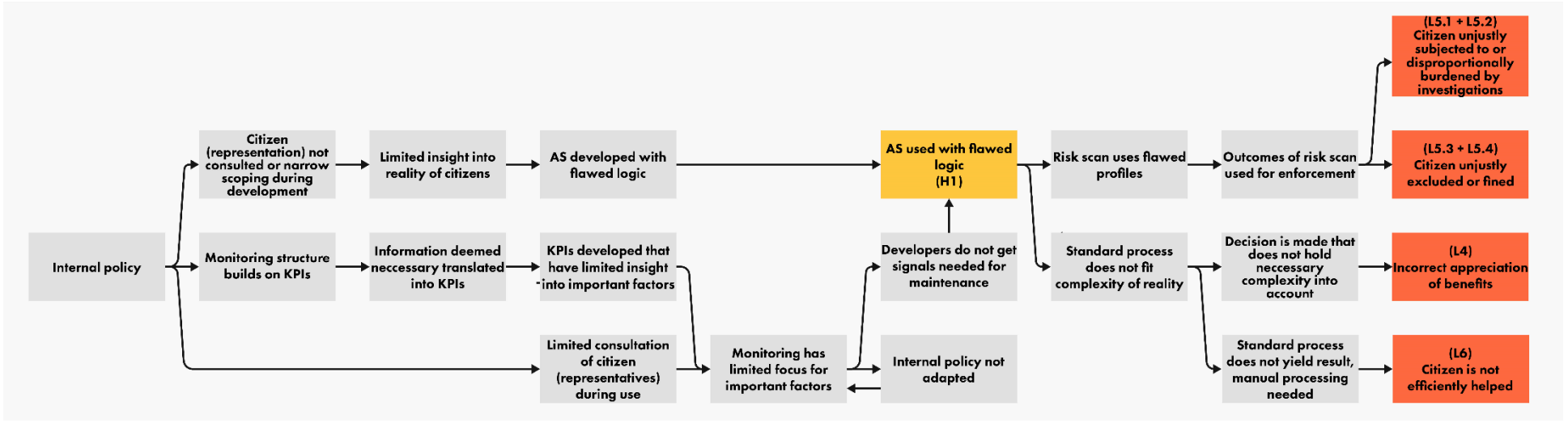

Figure 17 Causal Scenario 1: flawed logic at UWV

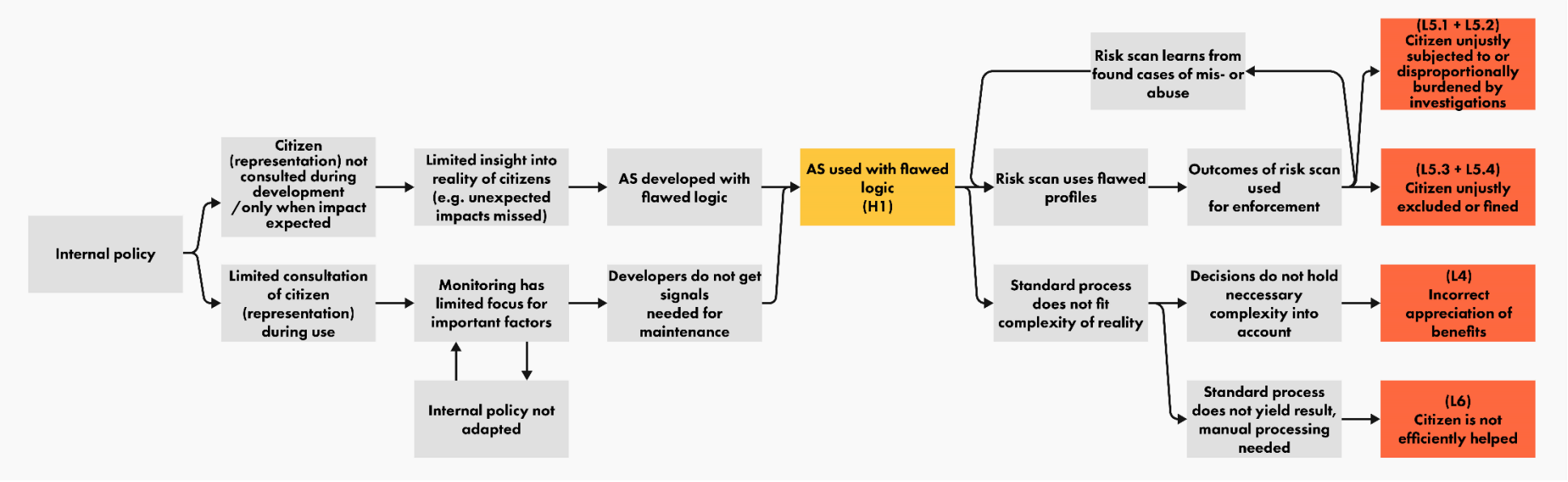

Figure 18 Causal Scenario 2: flawed logic at SVB

5.6.2 Causal scenario 3: Use of flawed information (H2 & L4, L5)

This causal scenario, displayed in Figure 19, pictures the effects of flawed information permeating through coupled systems. UWV and SVB make use of SUWInet, which allows them to attain data from other organisations. There can be errors present in this data, or it can be outdated. The information can also be present from an organisation that has a different use case for the information, which can mean definitions differ. This can mean the information might not be fit to use in the SUWI-organisations. This flawed information can then be used to (re)appreciate benefits, either automatically or not (H2). The appreciation of benefits based on flawed information can cause unjust use or incorrect appreciations (L2 & L4). Changes in information can also trigger investigations in to mis- or abuse (L5).

This causal scenario depicts how hazards can arise from causes that cross boundaries of organisations. This impacts the ability of UWV and SVB to oversee and control this hazard: they depend on other organisations. Within the execution of social security schemes, data sharing has increasingly become important. Several reports have pointed to the possible side-effects. Organisations have grown increasingly dependent on each other for their data and the quality thereof. Changes in one system can lead to consequences elsewhere, impacting traceability of information and related decisions and thereby transparency (Stichting Kafkabrigade, 2023; van Atteveldt et al., 2024). It brings about questions of control over information, and thus accountability over the information and use thereof. Definitions of information can differ between organisations, but this is not always visible for the organisation using the information (Widlak, 2022).

Solution directions

The solution directions this causal scenario inspires are interventions higher in the SCS. The first two causal scenarios asked for interventions between developers and citizens, citizens and executing employees, and eventually management within the two executing organisations, this causal scenario primarily asks for interventions placed in between organisations exchanging information.

The organisations, in order to prevent flawed information being used, need to be able to be in control of their processes and take responsibility of data used. This means they must be able to have some understanding of how the data was gathered and stored, so they can oversee currentness, quality, use cases et cetera. This asks for communication between organisations, to understand how other organisations handle their data. Beyond that, it increasingly asks for agreements between organisations that share data, to come to standards for definitions, data quality, storage and so on (e.g. also CIA-triad). This asks for discussions of higher management of these organisations, most likely with the involvement of relevant ministries (e.g. Social Affairs). Being able to have these discussions asks for mental models that include a good picture of what the data is used for within organisations and how exchanged data is gathered. Centralisation of AS development or data gathering has also been suggested in interviews (I1, I7, see also Stichting Kafkabrigade (2023)).

For this picture, another intervention might be relevant. That is, mapping interdependencies between different data streams. For instance, how changes in income lead to changes in entitlement in benefits, and whether these changes are automatic or not (triggers). This not only allows to see what data especially needs to be of high quality (i.e. data that many decisions rely on), but it also allows to understand complexity of the system, which in turn

allows for executing employees and citizens to be able to better understand how things come to be. It makes decisions and changes traceable and thereby aids motivation of decisions.

System-level constraints

General constraints were (paragraph 5.3.4);

- Data quality must be upheld;
- There must be ownership of data.

The possible interventions lead to following set of additional constraints, that can be seen as further specification of the general constraints;

- Organisation using data gathered and/or maintained by another organisation must have sufficient understanding of the data, i.e. how it is gathered, what definitions are used, how it is stored and what its quality is;
- Organisations sharing information on a structural basis must have common standards on definitions, quality, storage and handling of data;
- Organisations sharing information must have a sociotechnical mapping of connections within systems and related (automatic) influences between data, this must include the legal connections and how they are translated into technical systems;
- Decisions made on data that was shared between organisations must be traceable.

Input from interviews

Asides from mentioned literature, the Actor Analysis, and the previous paragraphs on the SCS and control actions, this scenarios and possible solution directions builds on insights from interviews. In particular, it relates to subthemes ‘interdependencies and data coupling not traceable for citizens’ and possible solution ‘centralisation of certain tasks’ (see Appendix E and Chapter 6).

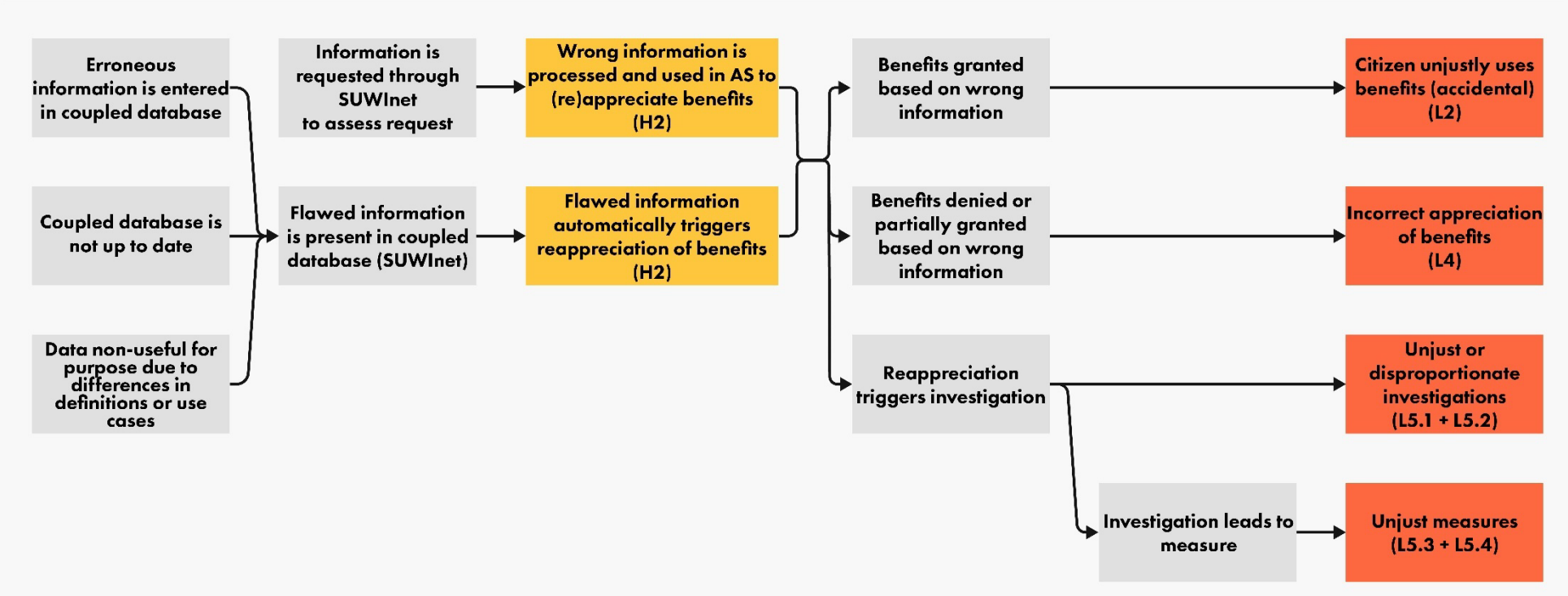

Figure 19 Causal Scenario 3: flawed information

5.6.3 Causal scenario 4: Citizen has flawed information (H3 & L7)

This causal scenario, pictured in Figure 20, shows how citizens can have little information on involvement of AS, and together with limited information in the supervisory landscape, this can lead to citizens being unable to exercise their rights. It combines all three SCSes.

Within the two executing organisations, the interconnection between data sources together with increased technical abilities leads to complex AS and effects that are difficult to oversee (as described under causal scenario 2). Together with the complex and interconnected social benefits schemes, this leads to limited motivation of decisions. That is, the motivation only gives a limited view on the involvement of AS and data decisions were based on. Internal policy can influence this motivation. For instance, fear of gaming the system can cause executing organisations to limit information they give about risk scans and information used therein. This causes citizens to be unaware of the role AS played in the decisions that were made about them. Because the source of information is not always clear due to data sharing, the citizen also might not know where to turn to in order to correct information. Here it also ties in with the general SCS: the algorithm register is not focused on a particular group, it offers little actionable information, and is incomplete, for instance due to its non-mandatory nature. So, the motivation given to citizens can offer too little information, and if the citizen tries to actively seek information, the algorithm register might not be able to bridge this gap.

In the general SCS, it is also visible that there is no central supervisory authority with formal powers. Supervisory organisations also suffer from a lack of means. This results in knowledge on AS use and AS hazards not being centrally created or retained. Even if cases arise, learning effects might not be captured. A citizen can choose not to act due to the little information they have, in which case they have been unable to exercise their rights. If they do take action and, for instance, complain to a supervisory organisation, they might not mention the AS because of their lack of knowledge. If needed knowledge is not present in the supervisory landscape, the involvement of AS might not be discovered or properly appreciated, in which case the citizen is also not able to properly exercise their rights. Furthermore, this means lessons are not captured, and knowledge is not added to the supervisory landscape.

This causal scenario thus shows that information given to citizens, but also present within the supervisory landscape can impact the ability to uncover the role of AS and therefore the ability to exercise rights. It also shows that if the role of AS is not uncovered, learning effects are not captured. This means harms by these AS can persist and governance is not able to learn and improve. For citizens, it can be especially hard to be able to attain the necessary information to, for instance, prove they were discriminated against: it requires them to provide proof of possible bias that they can hardly attain (Houtzager & Verbeek, 2022).

Solution directions

This latter causal scenario suggests solution directions within organisations, but also within the general supervisory landscape.

Ensuring that citizens have the right information and are able to execute their rights can be approached from two ways: top-down and bottom-up. In a bottom-up fashion, this asks for increased transparency from executing organisations to citizens. Primarily, this would mean including more information in motivation of decisions, if the motivation is given in the first place. That means the motivation has to include a proper description of the process and AS, so that citizens or their (legal) representatives can understand what transpired. This intervention

thus lays in the interaction between citizen and executing employee, but most likely requires internal policy and resulting instructions including standard descriptions and training. Gaming the system plays a role for organisations to limit disclosure, in particular when it comes to risk scans. It must be asked to what extent this is realistic risk, or rather an assumption based on views on citizens and culture. Furthermore, it must be asked what can still be disclosed even when gaming the system is taken into account, and whether information should be available to independent third parties if not disclosed to citizens themselves.

In a top-down fashion, it is important that knowledge is built and retained within supervisory organisations, in order to be able to understand or discover the role of AS in processes. Within the supervisory organisations, this thus asks for structural knowledge building and retention, which in turn requires additional means to be allocated by the responsible ministries. Furthermore, additional means for a central supervisor could strengthen knowledge retention, discussions between supervisory organisations, shared actions, and the ability to discover overarching trends and risks. This is currently captured to some extent in the role of the DCA.

An important component that is currently in development is the algorithm register. This register could inform both bottom-up and top-down streams. It currently is not complete, and for the AS that have been included, information is limited and often not actionable. It is important to ensure this register is more complete in order for it to contribute to safety. This thus involves not only ensuring more AS are included, but also that there is a minimal amount of information given about these AS. This most likely asks for policy by either internal affairs (for all government organisations) or social affairs (if only related to SUWI), in order to make the register mandatory at least to some extent. What additionally could help would be to offer the possibility for supervisory organisations to propose changes to the register. If they, in their supervisory activities find AS or information about AS that they deem to be important for the public, they could propose this to be included, adding pressure on organisations using the AS and ensuring a neutral eye to be on the register.

What is important to note in the proposed solution directions, is that the actions taken in the supervisory landscape should be structural and sustainable. Currently, there is a lot of attention for AS and their harms, but it is important to ensure supervisory means, actions and knowledge remains even if risks (seemingly) decrease. This is also described by Bokhorst et al. (2013) as the supervisory paradox: a call for more supervision when accidents happen, and less supervision when things are seemingly in order.

Furthermore, transparency without actionability will not sufficiently help prevent hazards. As Diakopoulos (2020) describes: “transparency is not sufficient to ensure algorithmic accountability [...], true accountability depends on actors that have the mandate and authority to act on transparency information in consequential ways’. This calls for information to be actionable. That is, the supervisory authorities and citizens must have means to act on the information about AS that is given, and vice versa, if the means are already present, the information given must be enough to use these means. Means could for instance be the ability to file complaints (citizens), or having the mandate to pressure those using the AS to improve processes (supervisors). The ability to use existing means could for instance be sufficient information to explain involvement of AS to a supervisor, or the ability for a supervisor to understand where in a process AS are used. For citizens, actionability could also mean ensuring motivation of decisions, or information provided in the algorithm register, includes what they can do with this information. It could also mean the algorithm register, or other information

regarding AS in use, is mentioned so citizens or their (legal) representation know to look for it in the first place.

System-level constraints

General constraints set under paragraph 5.3.4 were:

- There must be active transparency
- There must be passive transparency
- Information must be understandable for citizens

These possible interventions lead to the following additional constraints:

- Motivation of decisions must include a description of the process that lead to it, including the role of AS therein
- Non-disclosed information must be available for checks by an independent third party (e.g. basis for risk scans)
- Supervisory authorities must structurally invest means into knowledge on AS, and therefore sufficient means must be provided to them
- There must be legally defined rules on which AS need to be disclosed in a central register, and what information must be included about AS
- Information given on AS due to transparency requirements must be actionable

These constraints are additional to constraints set under paragraph 5.3.4. When it comes to transparency, these constraints can also be seen as further specification.

Input from interviews

Asides from mentioned literature, the Actor Analysis, and the previous paragraphs on the SCS and control actions, this scenarios and possible solution directions builds on insights from interviews. In particular, the causal scenario builds on the following six subthemes (see Appendix E and Chapter 6);

- Average citizen not aware of use of AS;
- Interdependencies and data coupling not traceable for citizens;
- Bottom-up signalling by citizens of AS involvement;
- Supervisory organisations do not always have direct knowledge of AS involvement;
- Knowledge not well documented and retained;
- Algorithm register is incomplete and not attuned to citizens' needs.

The possible solutions relate mainly to subthemes 'central and proactive supervision', 'independent supervision on risk scans if not disclosed', 'more complete algorithm register' and 'increased transparency, motivation of decisions and actions' (see Appendix E and Chapter 6).

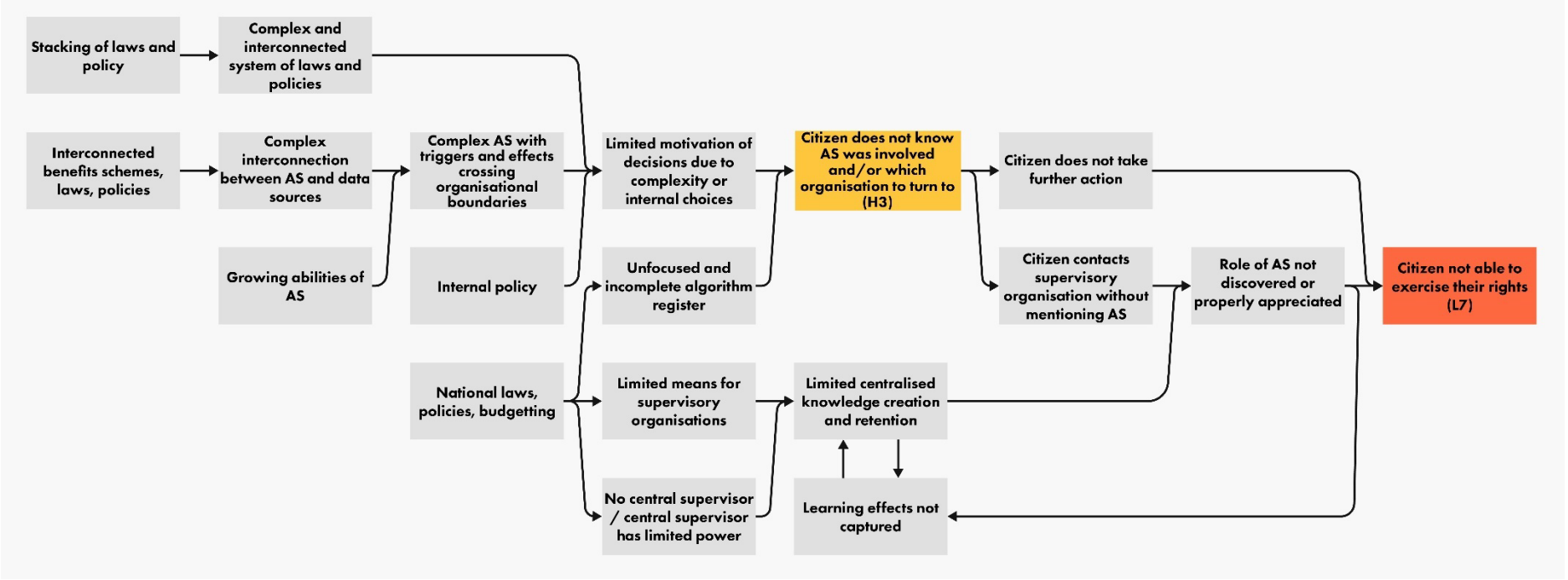

Figure 20 Causal scenario 4: flawed understanding

6. Empirical insights from interviews

This chapter describes the approach and results of the interviews that were elicited to learn from practice. First, the structure of the interviews will be discussed, after which follow the results. The interviews aimed to answer the following sub-question:

In practice, what can be learned about the governance of algorithmic hazards at executing agencies, and how can this be understood in the context of an Actor Analysis and Systems Safety Analysis?

6.1 Structure

Semi-structured interviews were held with eight respondents. The interviews had four themes. The Actor Analysis provided input for these interviews. The interview protocol and the additional focus and information used for each interview can be found under Appendix C. An overview of interviewed actors and related appendices can be found in Table 5. The four themes used were;

1. What does your organisation do and what is your role?
2. What do you see in your role?
3. What needs to happen within organisations?
4. What needs to happen within the wider supervisory landscape?

The first theme was meant as an introductory question, but also to gain further insight into what is currently happening, serving as validation and feedback for the Actor Analysis and STPA. The other themes are in addition to it, and meant to further understand general observations of respondents, and what could be done to improve governance. Interview reports can be found under Appendix D.

Table 5 Overview of interviewed organisations

Actor	Interview	Appendix
National Ombudsman	I1	D.1
Amnesty Netherlands	I2	D.2
UWV	I3	D.3
Court of Audit	I4	D.4
National IT Guild	I5	D.5
Netherlands Institute of Human Rights	I6	D.6
National Client Council	I7	D.7
Internal Affairs	I8	D.8

6.2 Subthemes

The last three themes (observations and governance improvements) were analysed to find overarching subthemes mentioned in multiple interviews. An overview of these themes can be found under Appendix A. What follows is a description of the different subthemes that were found.

6.2.1 Observations on governance of AS

When it comes to the use of AS in decisions and investigations that impact citizens, these citizens themselves are often unaware of AS involvement. Citizens have too little knowledge on what happens within organisations, and thus do not know AS were involved, unless told.

Citizens are also not aware of frequent data sharing and interdependencies in law and AS, and thus are unable to trace mistakes. The lack of knowledge of citizens is especially troubling, since signalling issues of AS at supervisory authorities primarily relies on citizens themselves. As a result, very few complaints explicitly mention AS involvement. Legal protection is bottom-up, but a citizen does not have sufficient information about organisations and only knows their own case, not all others.

When AS involvement is not mentioned by citizens, it can take a while for supervisory authorities to figure it out. They need the right knowledge in order to ask the right questions or make discoveries themselves. Knowledge is lacking, however, also within the departments using the AS, which provides challenges for supervision. Involvement of AS can sometimes be suspected through certain signs, such as decisions that are stalled or patterns of citizens being subjected to checks. For supervisory authorities, it is still new in practice. Even if certain knowledge is present, it is not always well retained. Knowledge gathered around a certain case or publication does not always stick. Furthermore, cumulative knowledge is not built. Cases and citizen stories can provide valuable insight into mistakes and best practices, but there have not been many cases that can provide needed knowledge.

There are also certain knowledge gaps specific to rights, values, and technical systems. Organisations have difficulty translating certain human rights and values to their digital systems in their context of use, and want to know what they mean in practice and how they can show compliancy. These rights are not unambiguously explained by an authoritative party, and knowledge in organisations can be insufficient. Technical employees can have difficulty recognising risks for human rights, and other issues of political and ethical nature. Organisations might have good technical monitoring, but do not always have a view on societal implications or plausibility of assumptions. Citizen groups are not always involved, even though the realities of citizens can be more complex than systems can account for. Furthermore, certain values and fundamental rights can be difficult to quantify and monitor.

Respondents are positive about recent actions, but also critical of them. The algorithm register is not attuned to citizens' needs and gathers more attention from journalists and academics. Citizens that need benefits might not be focused on such a register in reality, although it is ideally meant to inform them. The non-mandatory register is also incomplete, especially when it comes to systems that can possibly do harm, although it is slowly being filled. When it comes to laws, regulations, frameworks and guidelines, these can be a lot for organisations. They come from different areas, can overlap, and also overwhelm. The different current actions do not guarantee safety: even if guidelines and assessments such as IAMA and DPIA are followed, mistakes can be made. Ethical committees might look at ethical permissibility, but leave out certain other implications that might be important. A guarantee that mistakes will not be made, however, might not be possible. It is instead necessary to look at what is 'good enough'.

Organisations are aware that AS bring about risks, and want to do things right. However, it is not defined what 'good enough' is, leaving organisations wondering. Not knowing what is good and wrong causes insecurities for organisations. This is especially true in an environment where mistakes lead to social outrage, even if handled correctly. Attention for AS has increased massively, but problematising AS can steer away from core issues. Furthermore, backlash can impact willingness to share and thus ability to learn from mistakes.

The mentioned observations have been included in the analysis described in Chapter 5. The different observations have been used to support all four causal scenarios and thereby helped explain how hazards come to be and how they lead to losses. This supports the theoretical underpinning of these causal scenarios.

6.2.2 Governance improvements in general landscape

In the broader supervisory field, several possibilities for improvement of governance were mentioned. Central and proactive supervision needs to be organised. A central contact point can help citizens, but also supervisory authorities. Central supervision can help prevent human rights issues in a top down manner, which is important where bottom-up action will not be sufficient. DCA is mentioned not just as a coordinator, but also a supervisor, overseeing AS, and artefacts such as the algorithm register. Mandate and means are necessary here. When it comes to risk scans, there are reasons not to disclose all information. Although this should be done as much as possible, where it is not possible, it should be accessible to an independent organisation to ensure human rights are safeguarded.

For all AS, it is important knowledge is structurally developed and retained. This refers to technical knowledge within supervisory organisations and politics, but also knowledge on human rights and values within more technical teams. Different organisations can be brought together to discuss what is needed and to weigh different interests. It is important this knowledge is documented in order to capture learning effects. Such knowledge can help further specify existing frameworks and guidelines and inform policy and standards.

More centrally, there needs to be further specification of sufficient measures and application of human rights and values. Different contexts warrant different trade-offs, and further specification of assessing human rights issues can be helpful, either by a supervisory authority, in central guidelines, or in bottom-up discussions. Top-down unambiguity regarding human rights can be of importance. Furthermore, a consensus is needed as to what is seen as sufficient for an organisation to do. This ties in with the public and political discourse needed. Guarantees that mistakes will not be made cannot be given, and it is thus about deciding what is sufficient. When mistakes are made, however, it is important to prevent outrage but rather learn from mistakes. Concrete cases, but also jurisprudence, can be helpful to learn and further shape governance.

In order to enable citizens to take action, but also for supervisors, judges, journalism and so on, it is important that organisations are more transparent about use of AS. A fitting way to be transparent should be found. Centrally, there already is the register, which therefore could be a logical instrument. However, it might be necessary to make this mandatory.

Relation to STPA

The possible improvements have been included in the causal scenario of hazard 3, which was described under paragraph 5.6.3. This makes sense, as this causal scenario had specific focus on the supervisory landscape, as does this part of the interview analysis. In addition to outcomes of the STPA, the proposed measures in the interviews provide more precise description of what knowledge needs to be retained, i.e. documentation of trade-offs that were made. Furthermore, there is an additional intervention that was not described under STPA. That is specification of sufficient measures in order to decrease insecurity of organisations using AS. This specification should be done by supervisory authorities or political

actors. BZK is currently working towards this, in a cooperative way. Involvement of other actors, such as supervisors, might help bring this forward. It should, however, also be considered that involvement of citizen representation, could also be of importance. As Swierstra and Vermaas (2022) describe, 'where the assessment of harm can often be left to scientific experts, the assessment of good is intrinsically contested and political. This means that establishing what can count as 'good' should not be left to the ethical experts. Citizens need to be invited to join the discussion about what is the right direction, and what is then the right technology.' This hints towards political involvement where citizens' values are taken into account.

6.2.3 Governance improvements within SUWI

Within organisations there is also work to be done on transparency in order for citizens to be able to defend themselves and to understand why decisions were made. Governmental organisations should clearly communicate what systems they use to make decisions. Citizens also need to know why they are being subjected to investigations. Some respondents call for centralisation of certain tasks. For instance, a central point where data is stored and shared, or a central point where AS are developed and maintained for different executing organisations. This means citizens know where to go when issues arise with their data, but also that supervisory authorities know where to look in order to supervise AS. Connecting more to the world of citizens and including societal groups is also mentioned. Representatives of those affected, such as client councils, can be included to make suggestions, not just when mistakes were made but also proactively. This can also include other groups such as employers, unions, and so on. The reality citizens live in is complex and including societal groups can bring unforeseen realities into view.

Not all relevant factors can be quantified or foreseen, therefore it is important to have the right culture in which discussions take place with multidisciplinary knowledge present. This can help understand what certain values and rights mean in context and how they should be governed for. Connecting to citizens and having the right culture also allows to make integral trade-offs, where both opportunities and risks are considered and a decision is made whether use is wanted. For this it is thus necessary to consequences for citizens and society. Even if things are possible, we should wonder if we should want it. Furthermore, it is necessary to be able to oversee subsidiarity, for which measuring a baseline is needed. This can become difficult if AS change systems and society.

Asides from integral trade-offs, it is needed to integrally apply existing and upcoming law. The different existing laws, policies, but also frameworks have similar underlying purposes and duplicate components that can be applied. Some existing practices and questions present in the past still apply to AS, such as the use of risk profiles and motivation of decisions. The AI act will bring extra safeguards, but will not fully safeguard all that is necessary. Therefore, organisations should not wait for the act. Furthermore, it is necessary not to wait for exact definitions and applicability of the AI Act. The risk-based approach the act takes on can already be applied, even if the exact details are not known yet.

Relation to STPA

The possible solutions have been included in all causal scenarios under paragraph 5.6. In addition, the interviews mentioned centralised discussion including societal groups, centralisation of data gathering and the possibility of centralised development and use of AS.

The interviews also provided a more general intervention, that is: making integral trade-offs between benefits and possible risks, including effects on society, before deciding whether to use an AS. This intervention could take place within organisations that aim to use the AS, when it comes to specific purposes of AS, and should happen with a diverse group of experts (including citizen representation). Another way, is more centralised trade-offs about generic types or purposes of AS, where the goal is to decide whether they should be used or not. This should happen higher-up, with the involvement of politicians and supervisory authorities.

6.3 Additional lessons for STPA and AA

The interviews provided additional insights that add context to the research findings and mappings made in AA and STPA. The court of audit was often mentioned as part of the 'supervisory authorities', but the interview (I4) highlighted that the organisation is not part of this landscape, but an addition to it. The different interviews also highlighted contacts between organisations that could impossibly all be mapped, but might influence the case at hand. Examples are discussion platforms (e.g. 'raadbaak', I1), discussions on law proposals (I7), and cooperations when creating and applying frameworks (I4). Furthermore, different organisations highlighted direct contact with citizens, partly through campaigns that asked for citizens to contact organisations (I1, I2, I6, I7). Lastly, different organisations work on several themes that are part of their focus for a certain amount of time (I1, I2, I4, I6). This allows the organisations to focus on topics that they see risk in, which is necessary due to limited means. However, one could wonder what the influence of working within certain themes is on the sustainability of knowledge development, and how this relates to the 'supervisory paradox' (see 5.6.3).

The interview with UWV (I3) provided additional context to the development and use of risk scans within that organisation. UWV builds risk scans based on hypotheses only, not based on data-analyses. Their risks scans do not improve over time, but are relatively simple models that are 'frozen'. The enforcement division of UWV does not work with the outcomes of the scans, this stays within primary divisions. Different types of monitoring are used, namely continuous and more detailed periodical monitoring. Part of the periodical monitoring is the complaint intensity of decisions made based on the scans. Furthermore, the 70/30-rule allows not only help prevent overreliance, but also provides a continuous control and experiment group to monitor the AS. The policy surrounding risks scans that this was part of was created relatively recently, and as developments are fast, some parts of the policy have not been applied yet. Lastly, the respondent highlighted how risk scans are at the end of a ladder of interventions aimed to prevent misuse of social benefits. The risk scans, at the end of this chain, focus on a small group of citizens that do not respond to other interventions and are assumed to purposefully break rules. References to lack of willingness of UWV to disclose information on the risk scans should always be seen in this wider context of interventions.

The interviews also highlighted some important actors and components that were not included in analyses. The legal system in the Netherlands was deemed important but not included, although jurisprudence can be of importance in this case. Internal complaints procedures and auditing were also not included in the STPA-analysis. Legal aids of citizens, as legal representatives, were also left out of scope, even though several organisations deal with them instead of or in addition to citizens. BKWI, the organisation responsible for data exchanges in SUWI was also not specifically included as a separate organisation.

7. Actor Analysis and STPA

Chapter 3 provided initial considerations of the combination of Actor Analysis and STPA, focussed on the following sub-question:

What are similarities and differences between Actor Analysis and System-Theoretical Process Analysis, and what do these mean for combined usage of these methods in describing AS?

This chapter returns to reflect on the use of AA and STPA together based on these initial considerations.

7.1 Comparison of methods and consequences

The two methods were compared based on the aspects explained in Table 1 (Paragraph 3.2). A summary of this comparison together with the consequences for the outcomes of these methods can be found under Table 6.

7.1.1 Purpose of analysis

The purpose of the two analyses differ: Actor Analysis seeks to analyse a problem situation, whereas STPA seeks to identify (potential) problems. By using AA before STPA, there is a known problem statement as input for STPA, which serves as a basis for the analysis. AA posed an enriched problem statement, and STPA was used to explore what this problem meant in practice and how it can materialise into losses. AA set the stage and provided a knowledge base, STPA explored further how the posed problems could be understood.

7.1.2 Methods

STPA and AA have methodological differences, which translate into differences in scope and differences into the included elements. AA is inter-organisational, whereas STPA typically looks within one or few organisations surrounding a controlled process. While AA provided a solid structural basis for the general SCS, it only provided a baseline of knowledge for the individual SCSes of UWV and SVB. The difference in scope therefore impacts the ability of AA to inform parts of STPA that deal with what happens within an organisation. The general SCS can be seen as a new type of artefact that helped bridge the two methods. It does not include a controlled process and is therefore not a typical SCS. However, it did help bridge the two methods and widened the scope of STPA, giving ability to translate issues from an inter- to and intra-organisational view. As the problem at hand involves policy and supervisory issues in the wider inter-organisational landscape, this bridging is important.

The general SCS also builds on similarity in showing hierarchical relations, AA shows hierarchical formal relations, STPA shows hierarchy in control actions and feedback. This means STPA is able to build on AA, but also that relations between units can be compared. AA poses formal relations, STPA shows how they translate into control and feedback, and vice versa. For instance, the control and feedback to and from the algorithm register shows potentials for hazards, which in part can be derived back to the lack of a formal relation underlying this connection.

When it comes to the units, AA looks at actors, STPA looks more widely at controlled processes and the controllers in the SCS. Because of this, STPA is better able to show the sociotechnical context of the problem at hand. The flipside of the ability to show more context is that scoping is more difficult. A SCS can quickly become clouded, but including a variety of

units can be important to display context. This means framing the way problems are looked at is more important under STPA, which in turn increases the influence of choices of those executing the analysis.

Whether units are included or not also differs between the methods. AA includes actors, STPA includes units if they can control or are controlled. This means STPA can include more actors and also stakeholders, which again means it is better able to show context. It can include those with inability to act, and those who have no formal representation. In this research, including citizens and not just their representation increases the ability to explain how hazards of AS materialise and impact citizens and society. The inclusion of the press shows important context in how the problem is framed.

7.1.3 Values

Lastly, AA and STPA show underlying differences in how values are incorporated in the analyses. AA shows ability to include various values based on the actors included, and explain these values and their relations as the actors see them. Meanwhile, STPA only includes values relating to safety, and sees other values as hierarchically lower than safety. This, in practice, translates into an important shortcoming of STPA. Different values such as the double binds described under AA were only implicitly present and not explicitly used to understand how safety was not reached (i.e. why the problem was not yet solved). Furthermore, STPA did not include a description of differences in values actors find of importance and how they lead to issues or a lack of solutions.

When it comes to the viewpoint on safety, in particular, AA looks in a more descriptive and contextual manner, whereas STPA upholds a relatively prescriptive and definitive viewpoint. This difference shows an important shortcoming of STPA, as it can lack nuance and risks a negative bias. This difference is partly bridged by the fact that losses are based on stakeholders, meaning the notion of safety indirectly depends on stakeholders: losses are what stakeholders want to prevent, and safety is then the absence of these losses. However, still, STPA only looks into how losses can be prevented, not necessarily into how the differences in the conception of safety can be understood. This concerns both the definition of safety and the level actors find acceptable. Analysing what different relevant actors see as safe can contribute to understanding why losses can still occur or solutions have not been introduced yet.

Table 6 Comparison of Actor Analysis and STPA and consequences for the results

Aspect	Actor Analysis	System-Theoretical Process Analysis	Consequence
Purpose			
Purpose of analysis	Problem analysis: enriching a known initial problem statement	Problem identification: potential for safety problem assumed	AA provides rich problem situation for STPA and 'sets the stage'
Methods			
Typical scope	Inter-organisational: looks at different organisations in problem situation	Intra-organisational: looks within one or few organisations surrounding relevant controlled process	Difference in scope impacts initial focus of STPA and ability to inform analysis
Unit(s) of analysis	Actors: typically organisations or representation of individuals	Controllers and controlled processes (various)	STPA has better ability to show sociotechnical context, but can be difficult to scope and is therefore influenced more heavily by researchers' choices
Constellation of units	Hierarchy, formal relations	Hierarchy, control and feedback	Similarity creates information synergy, and ability to compare formal relations and translation in practice
Inclusion of units	Has capacity to influence decision-making or to act on decisions and their outcomes (actor)	Part of safety control structure: involved in the problem situation in practice (can control or is controlled)	STPA can include more actors and stakeholders and show those with inability to act
Values			
Inclusion and relevancy of value(s)	Actor-induced, value plurality: different values relevant to actors in problem situation	Safety-related, limited value plurality: values relevant to safety	Different values not used in STPA to understand how safety is not reached
Conception of values and their relations	Conception of different values and their relations through conceptions of relevant actors	Conception of different values through their influence on safety, relations are hierarchical, where safety is the superior value	Different in value conceptions and relations not used in STPA to understand how safety is not reached
View on safety	Descriptive, contextual: dependent on actor	Prescriptive, definitive: absence of losses	STPA lacks nuance in conception of safety, and risks a negative bias.

7.2 Synergy between methods

The use of AA and STPA together in this research brought forward synergy between the methods in how they can be used to describe sociotechnical systems. The comparison in paragraph 7.1 also brings forward several sidenotes for the use of the methods in combination.

Figure 21 shows the way in which AA, STPA, and the interviews provided value to each other. Actor Analysis provided a knowledge base by sketching the problem situation and uncovering information that was used to execute STPA and the interviews. For the interviews it provided an overview that was used to contact possible respondents, and knowledge gained about actors was used in preparation for the interviews. For STPA, it set the stage with a basic hierarchical structure and knowledge that could be further explored.

The interviews allowed to test assumptions made in the Actor Analysis and STPA. This is why Figure 21 shows circular arrows for AA and STPA: they were performed iteratively, taking new

knowledge from interviews into account. For STPA, the interviews provided an important source of knowledge. It helped scope the analysis, allowed to more easily decide which hazards to take into account, and how to construct the causal scenarios. Furthermore, it helped test assumptions made in constructing the SCSes, improving them, as was done with the Actor Analysis.

In this synergy, the methodological differences and similarities meant STPA was able to build on AA, bridging the most important gap by use of the general SCS. The difference in units and reason for inclusion meant STPA was able to show the sociotechnical context better, and also able to depict more stakeholders that are of importance. When it came to the differences in how the two methods treat values, however, there is work to be done to improve the synergy between the methods. The double binds that were found in the actor analysis provide a nuance that was not further explored in the STPA. There is room to bring the way values are treated closer together, and have more attention in STPA for value conflicts and different views on values, including safety.

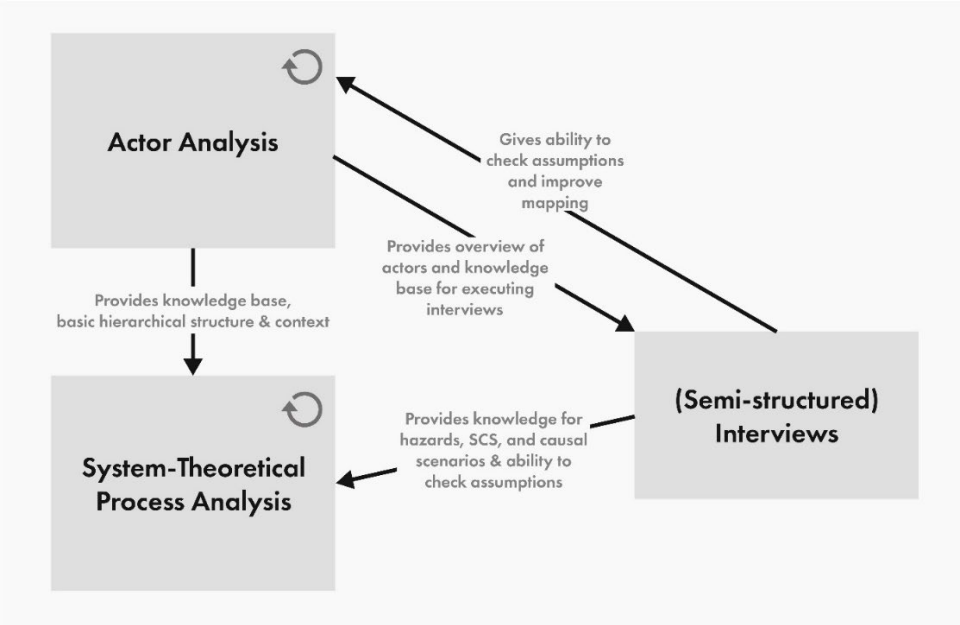

Figure 21 Use of AA, STPA and interviews together

8. Conclusion and discussion

This chapter describes the conclusions of this research, followed by limitations, a discussion and recommendations.

8.1 Conclusions

This research looked at the governance of algorithmic systems at two agencies executing the SUWI-law. Using a systems safety analysis together with an actor analysis and interviews with experts in the field, it aimed to answer the following main research question:

How can a systems safety perspective combined with an actor perspective be used to provide a sociotechnical description of algorithmic systems, their governance and possible hazards at agencies executing the SUWI-law?

In order to answer this main research questions, four sub-questions were used. The main conclusions on these sub-questions will now briefly be discussed.

8.1.1 Actor Analysis

How can an Actor Analysis be used to describe the actor field surrounding algorithmic systems at agencies executing the SUWI-law?

The Actor Analysis provided an overview of the actor field and relations therein, with several main observations. The coordinating supervisor, DCA, currently has no formal powers. Their ability to exert influence therefore currently rests solely on informal means. The Ministry of Social Affairs was seen as a non-dedicated actor, as there was little evidence to suggest they take action to govern AS at UWV and SVB. There is still potential for this actor to take action in the supervision of the two executing agencies. The analysis showed several groups of actors that were similar in dedication and power, resulting in a potential for cooperations, even though there might be differences in foci (e.g. a focus on discrimination, human dimension, privacy). These potentials have already largely been fulfilled in different existing cooperations. Executing agencies work together in the Manifest-group, different supervisory authorities work together in several discussion groups surrounding AI and algorithms, and there exist formal relations that ask for discussion between different political actors, and between different parties under the SUWI-law.

The Actor Analysis also provided insight into ambivalences in values that are strived for, both between and within different actors. These were described as ‘double binds’, and explain different goals or values actors strive for, but that can conflict and require trade-offs. The different double binds are governing AS versus governing by AS, carefulness versus efficiency, serving citizens versus taking strong action against fraud, and finding fitting solutions versus guaranteeing (legal) equality. Where the trade-off lands can depend on different factors, including the reigning political and societal climates. These double binds can help contextualise the issues use of AS in executing laws can cause. In all, the analysis provided important context to the problem situation, and a starting point for the other research questions.

8.1.2 System-Theoretical Process Analysis

How can a Systems Safety Analysis be used to describe algorithmic systems at agencies executing the SUWI-law, and the potential hazards of these systems?

The analysis focussed on seven losses for citizens and society that were explained as a result of three hazard categories: the use of flawed logic in AS, the use of flawed information in AS, and the possibility that citizens have a flawed understanding of the benefits system and the role of AS therein. STPA was used to make three mappings of the governance of AS. A new inter-organisational view, and one intra-organisational mapping of UWV and SVB each. These mappings, the Safety Control Structures, were used to analyse how hazards can arise and lead to losses. Causal scenarios were used to understand this in more details and make suggestions for improvements.

The analysis saw the use of flawed logic as a result of limited involvement of citizen representation, leading to a limited insight into citizens' realities being translated into AS and the monitoring thereof. The focus of UWV on KPIs increases this risk, whereas the self-learning nature of a new risk scan by SVB allows for the possibility of (runaway) feedback loops. The use of flawed information was related to the data exchange within the SUWI-system, leading to effects that cross the boundaries of individual organisations. Flawed data stemming from differences in data quality and use cases of data can theoretically be (automatically) exchanged and cause effects for citizens. The ability for citizens to have a flawed understanding of the benefits system and role of AS therein was related to complexity of laws and policies, leading to complex and coupled AS. Together with lacking motivation, this can lead to a citizen having a bad understanding of decisions involving AS. The lack of focus and completeness of the algorithm register does not currently solve this issue. This means a citizen will probably not be able to mention the involvement of AS when bringing up issues. The limited power that the central supervisor has, together with limited means for other supervisory organisations results in knowledge not being created and retained as well as it could be. This means that learning effects are not captured. It also means that supervisory authorities have limited ability to uncover the role of AS when citizens do not mention this, leading to the role of AS not being appreciated Based on this understanding of the possible harms that can stem from the use of AS in execution of the SUWI-law, several solution directions were posed. Initial, generic constraints, could be further specified and additional constraints could be stated. These can also be related to the mapping of the AS in their context.

Most importantly, STPA introduced a new way of looking at AS in their sociotechnical context. The analysis had the ability to show AS in the process they were part of, and relations to both technical and social components, such as humans, roles, organisations, and so on. The introduction of hazards, constraints, and losses, the depiction of Safety Control Structures, and the construction of causal scenarios provided a lens through which the safety of these systems could be understood and discussed. The general SCS provided a wider scope to STPA, showing not only what happens within, but also between organisations. The inclusion of knowledge from reporting and interviews showed that STPA was able to place real-life observations in a structured analysis.

8.1.3 Empirical insights from interviews

In practice, what can be learned about the governance of algorithmic hazards at executing agencies, and how can this be understood in the context of an Actor Analysis and Systems Safety Analysis?

The eight interviews with respondents at different involved organisations provided insights into observations on governance of AS, and governance improvements both within the wider landscape and within SUWI. The Actor Analysis proved to be an excellent source for selecting organisations and conducting the interviews. Subthemes were found under these main themes by use of coding. When it comes to observations on the governance of AS, findings point towards a lack of knowledge of both citizens and supervisory authorities. This poses risks, as problems will not be uncovered bottom-up, nor top-down. Organisations have difficulty relating human rights issues to technical systems they use, and might not have a good view on non-technical issues, such as complexity citizens find themselves in. Organisations find themselves in insecure environment, where they do not know what sufficient actions to govern AS are, yet where they receive backlash when issues arise, even if handled correctly. Interviewees were generally positive about current governance actions, but note the algorithm register does not fulfil its promised potential and that there is an overwhelming amount of frameworks and guidelines that overlap and can overwhelm.

In the general landscape, the interviewees offered different possibilities for improvement of governance. There needs to be central and proactive supervision over AS and their risks, and independent supervision over risk scans, especially when details are not openly disclosed. Knowledge needs to be developed and retained, also centrally. Political actors need to better specify what interventions are sufficient, and what different human rights and values mean for executing organisations and their AS. This also involves a different discourse, where absence of a complete guarantee of safety is accepted. There is a need for concrete cases and jurisprudence to learn from, and steps need to be undertaken to make the algorithm register more complete, possibly involving making it mandatory.

Within the SUWI-system, interviewees offered possible governance improvements. Organisations using AS need to work on transparency and motivation of decisions. They need to connect more with citizens and societal groups, in order to understand complex realities and make better trade-offs. Trade-offs need to be made integrally, looking at goals, benefits, but also possible harms and drawbacks. A culture where discussions can take place is important, and multidisciplinary discussions are valuable. Existing and upcoming laws and regulations need to be applied integrally, looking beyond the precise details and applicability and asking what the intentions of the laws and policies can provide. Centralising tasks, such as the gathering and storage of data and the development, of AS can prevent harms.

The interviews also provided context and improvements for the Actor Analysis and STPA. This includes formal and informal contacts between organisations and the fact that organisations work in themes for certain amounts of time, possibly impacting knowledge retention. Especially when it came to risk scans of UWV, the interview provided additional context, mainly the use of complaints in monitoring, and the hypothesis-based and non-self-learning nature of the risk scans. The interviews gave insight into components that were deemed important but not taken into account in the AA and STPA. Lastly, the interviews were used to execute STPA.

8.1.4 Two analyses: AA and STPA

What are similarities and differences between Actor Analysis and System-Theoretical Process Analysis, and what do these mean for combined usage of these methods in describing AS?

This research used an Actor Analysis and System-Theoretical Process Analysis in sequence. The use of methods and structured comparison of these analyses (Table 6) brought forward synergy in the purpose and methods used. AA as a problem analysis methods, provided a useful problem sketch as a starting point for STPA. In order to bridge the gap between AA, with an inter-organisational scope, and STPA, with an inter-organisational scope, a Safety Control Structure was added as a new artefact that depicted inter-organisational focus in STPA. This improved the explanatory value of STPA. The hierarchical nature of formal relations and components in the SCS meant the two methods aligned and AA could easily form a basis for STPA.

STPA includes more than just actors, and thus was better able to show the sociotechnical context in which AS are used and can lead to harms. STPA also includes more actors and stakeholders, and therefore is better able to show all those who are of importance, including those who do not have formal representation, or who are unable to act, such as citizens. However, as STPA includes more components, it requires more rigorous scoping. In scoping, the interpretation of the researcher becomes more apparent, making the analysis more subjective. The analysis done in this research, therefore, should be seen as an example of how AS and their harms can be conceptualised using STPA

Value differences between the two methods includes which values were included and how they were conceptualised. AA was able to include more values, as well as conceptualise and relate them more nuanced than STPA. As STPA upholds a more definitive notion of what safety means, it risks a negative bias. Although AA uncovered double binds, different value conflicts, they were not explicitly used in STPA. There is room to improve the synergy between the methods by bringing the way values are treated closer together, and have more attention in STPA for value conflicts and different views on values, including safety.

The use of AA and STPA in sequence, together with semi-structured interviews, gave the ability to show how the methods can be used to make a socio-technical description of AS and their harms with empirical basis. This can be found in Figure 21. AA can be used as input for interviews and STPA. The interviews can help improve AA, as well as guide the execution of STPA and validation of assumptions therein.

8.2 Limitations

8.2.1 Scope

This research focussed on the AS in use to execute the SUWI-law. The scope on SUWI as a whole meant that several AS were included, which in reality are used for separate social benefit schemes, each complex in their own right. Although several characteristics of processes and the AS used in them were taken into account, the discussion of all AS and processes stayed quite high-level. This limits the level of detail the analysis was able to provide, and means it possibly missed important context and specificity. Some extra notions were given regarding risk scans, but in general all AS were analysed in a similar way. Additional attention to the precise workings of AS and safety considerations different types of AS come with might have increased the ability of the analyses to provide insights for safety.

8.2.2 Methods

Actor Analysis

The actor analysis was based on a document analysis, looking into published documents to uncover information about actors and their relations. For some actors, it proved difficult to find information regarding the problem situation. For instance, the amount of documents on the Ministry of Social Affairs was low. This resulted in them being considered a non-dedicated actor. That means that the availability of information was seen as a proxy for dedication of an actor. It was found, however, that a lot more happens behind the scenes than is reported on.

There was also little information to be found on the different cooperations between actors. But even for those that could be researched to some extent, it was difficult to place these cooperations in the Actor Analysis. Even if a cooperation is to somewhat formalised, it does not constitute a hierarchical relation, neither does it a representation or membership relationship. The cooperations were placed within the analysis by describing them and adding elements to the formal chart, but their role and description perhaps does not reflect their importance in reality.

STPA

Similar to AA, it is difficult in STPA to display informal relations. Even if cooperations are known to happen, they need to be translated into control actions and feedback, for which there was little information available. Furthermore, including many cooperations can cloud an SCS. This is thus both a knowledge problem and a limitation of the method. Safety Control Structures used in STPA can quickly become hard to read, but by limiting what is included, context is missed. Significant scoping was needed to keep the analysis understandable.

STPA in this research project was based on documents and reports, as well as the Actor Analysis and the interviews. Although this provided a basis for the Safety Control Structures, these diagrams still involved significant choices by the researcher, including assumptions on relations between components. Combined with the need to scope for understandability, the SCSes and related causal scenarios relied heavily on interpretation by the researcher, allowing for subjectivity. The lack of information and assumptions made limit the validity of this research. It asks for increased involvement of organisations in order to be able to make a more accurate representation of reality. This was done to some extent in the interviews. In the end, however, although STPA showed the ability to conceptualise different observations in relation to each other, but stays mostly a theoretical understanding of how hazards can exist.

Actor Analysis and STPA

Although AA and STPA were combined in a structured way, reaping benefits of similarities and differences, the combination of methods mostly built on similarities and differences in purpose and methods. More could have been done not only to use AA as a methodological and knowledge basis, but also include values in a more structured way, thereby increasing synergy between methods and increasing the explanatory value of STPA. Although AA highlighted different values at play, they were not explicitly used in STPA.

Interviews

The interviews allowed to validate assumptions and add context to findings. Several organisations that were analysed during the Actor Analysis could have added additional value to the research and increased its validity, but were not interviewed. Within the ministry of social affairs, it was hard to find an expert that had knowledge on AS within SUWI. SVB was approached, but not available for interviews. The attempt to find more respondents within UWV also turned back with no result. Although the expert certainly provided valuable knowledge, it was solely focussed on risk scans. In general, it can be said that the amount of respondents per organisation and the amount of organisations interviewed can be seen as limitations, where an increase could have added to the validity of STPA. It can, however, also be noted that there is a lot of work ongoing when it comes to AS in government. This work is done by relatively small teams, which together with an increase in attention for AS, impacts their availability.

8.3 Discussion and main contributions

This research started out by posing that Algorithms and Artificial Intelligence have become buzzwords. The 'buzz' has seemingly expanded in recent years to include the harms that involve AS, and the need to take actions against it. This increase in attention and actions has its up- and downsides, that were implicitly present throughout this report.

Although such a buzz might bring attention to important problems, it must be wondered if reporting on AS and scandals gets to the core of issues. It paints a damning picture of AS as risky technical artefacts, leaving out the nuance that AS require. In reality, AS are used in a complex environment, that is multi-actor and socio-technical. In this context, there exist different viewpoints, existing political and legal regimes, each with their own complexities. The double binds that exist remain, even if political and societal environments change. Increased focus on the human dimension does not mean all notions of efficiency can be forgotten, for instance. The attention for governing algorithms leaves out the fact that executing agencies still need to govern by algorithms. The use of algorithms still happens within a context of complex social security laws and viewpoints that consider humans as possible fraudsters. Those viewpoints do not disappear overnight. The Actor Analysis and STPA provided a way to display the different structures that exist, including differences between actors and a more comprehensive view of the sociotechnical safety structures that surround AS.

What STPA also provided, was the ability to judge if and how different actions meaningfully lead to safety. The actions, including the algorithm register and different frameworks and guidelines, can be seen as positive developments that move towards safer AS. However, as part of a 'buzz', some interventions do not meaningfully add to the safety structures that exist. Some of the actions undertaken can perhaps be seen as solutions looking for problems they are to solve, produced with good intentions, but not enough understanding of the exact issues. The algorithm register, for instance, solves issues for research projects such as this one, but does not provide actionable transparency for citizens, which it intended to do. Different frameworks and guidelines overlap and overwhelm, and do not help executing agencies sufficiently. Meanwhile, what is good (enough) is still unclear for executing agencies, but not a question they should answer. The specification of values deemed important and application of human rights to AS asks for action by political actors. Furthermore, the specification of what is wrong is easier than of what is good, which is inherently political. Lastly, additional focus by existing supervisors and the instalment of DCA as a new coordinating supervisor are important steps, but they are meaningless without sufficient means.

The different actions and increased attention, even with their shortcomings, are positive developments, However, as part of a 'buzz', it must be wondered what will happen when the buzz dies down. Supervision over and knowledge of AS in the social domain needs to be structural and sustainable in a digitalised society wherein AS have become important components of execution and enforcement of social security laws. It is important to beware of the supervisory paradox, and ensure accidents are prevented instead of becoming the cause for a new call for increased supervision. At the same time, well-organised supervision still requires sensitivity from political actors. SyRI passed the judgement of political actors even though supervisory agencies had critiqued the system (Wieringa, 2023). Similarly, political actors missed signs of the ongoing childcare benefits scandal, as well as possibilities to address it (van Atteveldt et al., 2024). Even the best report does not help if no one reads it. Sociotechnical

views can help address how harms of AS are not a purely technical problem that can be solved by technical engineers, highlighting the role of managers and politicians. Not only can actions by political actors lead to executing agencies using AS in the first place, incentives and means given by these actors trickle down through the safety control structure to influence use of AS and the control thereover within executing agencies.

What these different notes on the 'buzz' show is how important it is to look beyond impulses and hypes, but try and understand the problem at hand and design meaningful interventions that fit this problem within its context. In complex socio-technical systems, understanding the issue can be hard, as can designing interventions, if not harder. Not only are causes never purely technical or social, neither are solutions. STPA and AA allowed to draw the socio-technical problem at hand. It allowed to understand where issues can arise, and how current actions lead to meaningful safety. Furthermore, it allowed to provide possible constraints and interventions, and to immediately place these within the existing control structure. The interviews helped provide some bridging between this theoretical understanding and the practice of governance of AS at the executing agencies. Together, it provided insights that ask for changes to all levels of the safety control structure: better involvement and abilities for citizens, changes to development, monitoring and motivation by executing agencies, increased discussion and agreement between executing agencies regarding data exchanges, better knowledge development and retention by supervisory authorities, means for these supervisors, and increased guidance, attention, and involvement of political actors.

The analyses provided a way to describe AS and their harms that can be used to come to shared understanding of what might go wrong and what can be done. As mentioned, these analyses built on interpretation and skills of the researcher, allowing subjectivity in the analyses. Although an important first step in coming to a shared understanding, work is to be done to build understanding together with researchers and practitioners, in order to create sociotechnical mappings with further theoretical and empirical underpinnings, which can bring the governance of AS forwards in a meaningful way.

To conclude, the main contributions of this research are the proposal it makes for a sociotechnical description of AS and their harms. It showed how this can be constructed using an Actor Analysis, semi-structured interviews, and a System-Theoretical Process Analysis, describing how these methods work together. In this exercise, it introduced concepts of STPA to describe AS and their harms: hazards, losses, constraints, understood through Safety Control Structures and Causal Scenarios. Moreover, the use of Actor Analysis in sequence with STPA showed how an inter-organisational view can be included in STPA, and pointed towards the possibility to add further nuance by using value descriptions discovered through AA in STPA.

8.4 Recommendations

The research findings allow for recommendations to be made for a central supervisor, policy makers, and future research.

8.4.1 Recommendations for central supervisor

For a central supervisor, it is important to take a role in ensuring central knowledge creation. This happens by documenting cases, developments, and bringing supervisory authorities and others together. DCA is already working on their coordinating role. It is important this knowledge is documented structurally, so it can provide lessons for governance of AS and top-down supervision. DCA can work on documenting knowledge centrally.

Work is being done on improving transparency surrounding AS. However, this research stated transparency must be meaningful and actionable. Bottom-up, it is important to see how citizens can be guided towards relevant authorities and how these authorities can pick up on signals they receive. Ensuring citizens can actually act on information provided to them by notifying relevant authorities makes transparency actionable. DCA is working on combining knowledge of contact points for citizens, but it is also important to uncover harms that are not brough forward by citizens themselves. Part of this lies within the bringing together of top-down supervision and uncovering certain patterns that may signal involvement of AS. But, in addition, it also requires a solid role for citizen (representation), in order to find out how and where hazards materialise. Only then can meaningful safety be provided.

Furthermore, it is important to keep an eye on different types of systems. Risk scans have been the focus of reporting and provide apparent risks for human rights, such as discrimination. However, decision-support systems and ADM-systems can also cause direct and indirect harms to society, even if these harm are less apparent or deemed less serious. The same goes for the data exchanges between organisations. It is important to keep an eye on these systems and couplings and their possible harms. DCA can do this through their risk reports, and in discussions they have with organisations using AS.

A focus on AS within their context is of outmost importance, in order to understand the full impact AS can have. This asks for DCA to have an eye for this wider, sociotechnical context. This involves working on a shared understanding of AS in their sociotechnical context, as well as a shared understanding of the hazards stemming from AS. This research and the use of AA and STPA can help this shared understanding forward. Understanding AS in their context can also help understand where issues arise, and thus what transparency should look like and how this can be made actionable (e.g. where do issues show up, how can these be communicated to citizens, and where should citizens be directed for help).

Although DCA can work on building a shared understanding of AS and their risks, push for actionable transparency, and help structurally document knowledge, their current lack of formal powers poses threats to their ability to do so. This asks for actions by policy makers, for instance to enable DCA to require information, be involved in discussions surrounding AS, where both currently rest on cooperation and goodwill. Furthermore, in order to shape meaningful transparency and to map systems and their harms, sufficient means will need to be provided to supervisory authorities by responsible policy makers and political actors.

8.4.2 Recommendations for policy makers

As mentioned, policy makers and political actors have to take actions in order to ensure (central) supervisory authorities can fulfil their roles. Policy makers also have an important role in ensuring supervision is sustainable, in order to prevent the supervisory paradox. Policy makers, as political actors, should also explicitly take their role in explaining what good entails, and thus what executing agencies should aim to strive for. The ministry of Internal Affairs is currently working on bringing different frameworks and guidelines together, which is important and useful to executing agencies. However, even bringing these different artefacts together requires specific considerations of what good means, not just what is not good.

Policy makers should push for further use of the algorithm register, mandatory or not, and provide further standards on what needs to be in it. This requires further specification of target groups and problems it aims to solve. Policy makers should also push for increased use of standards in data exchanges. Furthermore, it is important they take action in analysing the complexities of social security laws and interrelation between policies, and thereby also the complexity and interrelation between AS that are used. It is important that policy makers take action, but it is also important to prevent solutions looking for problems. Therefore, it is important to analyse where problems lay and how solutions can meaningfully lead to safety. This too, requires socio-technical analyses, such as STPA.

On the longer term, policy makers and political actors should consider how social security policies and the role of AS therein should take shape. Several political parties strive to get rid of the benefits system, or severely decrease the complexity thereof. It is important that the role of AS in the future is considered, whether they are used within less complex social security policies or otherwise. Political actors and policy makers should also consider the viewpoints on citizens and fraud within social security policies. There is increased attention for the human dimension, but these viewpoints need to materialise into changes into policies, but also changes into how citizens are viewed within executing agencies and government, especially when it comes to enforcement of rules. It requires specific consideration of how the government wishes to relate to citizens, including the role of AS therein.

8.4.3 Recommendations for future research

STPA can provide valuable insights into mapping of systems and hazards, as was shown in this research. The analyses in this research relied on assumptions in understanding what hazards are relevant, what components are relevant, how they relate to each other, and how hazards come to be and materialise. To increase added value, validity, and the use for moving towards meaningful safety, it is important future research seeks to involve organisations more in these types of analyses. Ideally, STPA is carried out together with executing agencies it concerns. Involving organisations using AS in analyses such as done in this research can help come to a shared understanding of AS and their risks.

This research made use of the methodological differences and similarities between AA and STPA to provide a valuable way to describe AS in a sociotechnical context. However, differences in how the methods treat values were not explicitly taken into account. The ‘double binds’ are an example of this. Future research should aim to further include the values and conceptions thereof identified under AA into STPA. Discussing different values, value conceptions, and value conflicts in STPA can increase the value of the analysis, help explain why hazards arise or solutions fail, and help come to a shared understanding of what safety is.

References

Algemene Rekenkamer. (2021a). *Aandacht voor Algoritmes*. https://www.rekenkamer.nl/publicaties/rapporten/2021/01/26/aandacht-voor-algoritmes

Algemene Rekenkamer. (2021b). *Vertrouwen in verantwoording: Strategie 2021-2025*. https://www.rekenkamer.nl/over-de-algemene-rekenkamer/publicaties/publicaties/2021/01/25/vertrouwen-in-verantwoording-strategie---2021-2025

Algemene Rekenkamer. (2022). *Algoritmes getoetst: de inzet van 9 algoritmes bij de overheid*. https://www.rekenkamer.nl/publicaties/rapporten/2022/05/18/algoritmes-getoetst

Algemene Rekenkamer. (n.d.-a). *Over de Algemene Rekenkamer*. Retrieved January 15, 2024 from https://www.rekenkamer.nl/over-de-algemene-rekenkamer

Algemene Rekenkamer. (n.d.-b). *Over het toetsingskader algoritmes*. Retrieved December 13th, 2023 from https://www.rekenkamer.nl/onderwerpen/algoritmes/algoritmes-toetsingskader/over-het-kader

Amnesty International. (2021). *Xenofobe Machines: discriminatie door ongereguleerd gebruik van algoritmen in het Nederlandse toeslagenschandaal* (EUR 35/4686/2021). https://www.amnesty.nl/content/uploads/2021/10/NL_Xenophobe-Machines_Amnesty-International_EUR-35_4686_2021.pdf

Amnesty International. (2023a). *Algoritme-'waakhond' dreigt een papieren tijger te worden*. Retrieved May 6th, 2023 from https://www.amnesty.nl/actueel/algoritme-waakhond-dreigt-een-papieren-tijger-te-worden

Amnesty International. (2023b). *Algoritmebeleid kan nieuw toeslagenschandaal niet voorkomen: mensen in Nederland nog steeds onvoldoende beschermd tegen discriminerende risicomodellen*. https://www.amnesty.nl/content/uploads/2023/12/AINL-2023-briefing-algoritmebeleid.pdf?x78975

Amnesty International. (2023c). *Amnesty International Afdeling Nederland Jaarverslag 2022*. https://www.amnesty.nl/jaarverslag/jaarverslag-2022

Autoriteit Consument en Markt. (2021). *Nederlandse toezichthouders versterken toezicht op digitale activiteiten door meer samenwerking*. Retrieved January 20th, 2024 from https://www.acm.nl/nl/publicaties/nederlandse-toezichthouders-versterken-toezicht-op-digitale-activiteiten-door-meer-samenwerking

Autoriteit Persoonsgegevens. (2022). *AP Inzet Artificial Intelligence Act*. https://www.autoriteitpersoonsgegevens.nl/sites/default/files/atoms/files/ap_inzet_ai_act.pdf

Autoriteit Persoonsgegevens. (2023a). *Jaarverslag 2022*. https://www.autoriteitpersoonsgegevens.nl/documenten/ap-jaarverslag-2022

Autoriteit Persoonsgegevens. (2023b). *Rapportage Algoritmerisico's Nederland (RAN) najaar 2023*

https://www.autoriteitpersoonsgegevens.nl/documenten/rapportage-ai-algoritmerisicos-nederland-ran-najaar-2023

Autoriteit Persoonsgegevens. (2023c). *Rapportage Algoritmerisico's Nederland (RAN) voorjaar 2023*. https://www.autoriteitpersoonsgegevens.nl/documenten/rapportage-algoritmerisicos-nederland-ran-voorjaar-2023

Autoriteit Persoonsgegevens. (2024). *Werkagenda coördinerend algoritmetoezicht in 2024*. https://www.autoriteitpersoonsgegevens.nl/documenten/werkagenda-coordinerend-algoritmetoezicht-2024

Autoriteit Persoonsgegevens. (n.d.). *Directie Coördinatie Algoritmes (DCA)*. Retrieved January 15, 2023 from https://www.autoriteitpersoonsgegevens.nl/themas/algoritmes-ai/coordinatie-toezicht-algoritmes-ai/directie-coordinatie-algoritmes-dca

Aven, T. (2022). A risk science perspective on the discussion concerning Safety I, Safety II, and Safety III. *Reliability Engineering & System Safety, 217*(January 2022). https://doi.org/10.1016/j.ress.2021.108077

Balayn, A., & Gürses, S. (2021). *Beyond Debiasing: Regulating AI and its inequalities*. https://edri.org/our-work/if-ai-is-the-problem-is-debiasing-the-solution/

Batini, C., Cappiello, C., Francalanci, C., & Maurino, A. (2009). Methodologies for Data Quality Assessment and Improvement. *ACM Computing Surveys 41*(3), 1-52. https://dl.acm.org/doi/10.1145/1541880.1541883

Bauer, J., & Herder, P. (2009). Designing socio-technical systems. In D. Gabbay, A. Meijers, P. Thagard, & J. Woods (Eds.), *Handbook of the Philosophy of Science: Handbook Philosophy of Technology and Engineering Sciences* (pp. 601-631). Elsevier. https://edisciplinas.usp.br/pluginfile.php/7503758/mod_folder/content/0/Designing-Socio-Technical-Sys_2009_Philosophy-of-Technology-and-Engineering-.pdf?forcedownload=1

BBB. (2023). *Iedere dag BBBeter: Van Vertrouwenscrisis naar Noaberstaat*. https://boerburgerbeweging.nl/verkiezingsprogramma/

Berg, A., Boot, T., Corrà, A., Kemme, M., Langeveld, M., Rooijers, S., Rotteveel, P., & Soldaat, M. (2019). *Onder de motorkap: analyse Garage de bedoeling: september 2016 tot heden*. Sociale Verzekeringsbank. https://www.svb.nl/nl/over-de-svb/wie-zijn-we/garage-de-bedoeling

Bits of Freedom. (n.d.-a). *Over ons*. Retrieved January 24, 2024 from https://www.bitsoffreedom.nl/over-ons/

Bits of Freedom. (n.d.-b). *Wat wij doen*. Retrieved January 24, 2024 from https://www.bitsoffreedom.nl/wat-wij-doen/

Bokhorst, A. M., Faddegon, K. J., de Goede, P. J. M., Knottnerus, J. A., & Welp, P. (2013). *Toezien op publieke belangen: naar een verruimd perspectief op rijkstoezicht* (e-ISBN 978 90 4852 211 8).

Brazier, F., Langen, P. v., Lukosch, S., & Vingerhoeds, R. (2018). Complex Systems: Design, engineering, governance. In H. Bakker & J. d. Kleijn (Eds.), *Projects and People: Mastering Success* (pp. 35-60).

Business.gov.nl. (n.d.). *About UWV (Employee Insurance Agency)*. Retrieved December 14th, 2023 from https://business.gov.nl/contact/about-uwv/

CDA. (2023). *Recht doen. Een hoopvolle agenda voor heel Nederland: verkiezingsprogramma 2023-2027*. https://www.cda.nl/verkiezingsprogramma

Client Council UWV. (n.d.-a). *Wat doet de cliëntenraad?* Retrieved February 1st, 2024 from https://www.clientenraad-uwv.nl/wat-doet-de-clientenraad

Client Council UWV. (n.d.-b). *Wat is de cliëntenraad?* Retrieved February 1st, 2024 from https://www.clientenraad-uwv.nl/over-ons

College voor de Rechten van de Mens. (2021). *Discriminatie door risicoprofielen: een mensenrechtelijk toetsenkader*. https://publicaties.mensenrechten.nl/publicatie/61a734e65d726f72c45f9dce

College voor de Rechten van de Mens. (2023). *In alle openheid: transparant algoritmegebruik door de overheid*. https://publicaties.mensenrechten.nl/publicatie/bf15558a-1b17-43d7-a60e-df9ff8847491

College voor de Rechten van de Mens. (n.d.). *Digitalisering*. Retrieved 15 January 2023 from mensenrechten.nl/themas/digitalisering

D66. (2023). *Verkiezingsprogramma D66 2023-2027: Nieuwe energie voor Nederland.* https://d66.nl/verkiezingsprogramma/

Delfos, J., Zuiderwijk, A., van Cranenburgh, A., Chorus, S., & Dobbe, R. ([Under Review]). *Integral system safety for machine learning in the public sector: an empirical account*.

DENK. (2023). *Nu is het moment: verkiezingsprogramma 2023-2027*. https://www.bewegingdenk.nl/verkiezingsprogramma/

Diakopoulos, N. (2020). Transparency. In M. D. Dubber, F. Pasquale, & S. Das (Eds.), *The Oxford Handbook of Ethics of AI* (pp. 197-213). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780190067397.013.11

Digitale Overheid. (2023). *Denk mee over het Algoritmekader*. https://www.digitaleoverheid.nl/nieuws/denk-mee-over-het-algoritmekader/

Dobbe, R. (2022). System Safety and Artificial Intelligence. https://doi.org/10.48550/arXiv.2202.09292

Elish, M. C. (2019). Moral Crumple Zones: Cautionary Tales in Human-Robot Interaction. *Engaging Science, Technology, and Society, 5 (2019)*, 40-60. https://estsjournal.org/index.php/ests/article/view/260

Enserink, B., Bots, P., van Daalen, E., Hermans, L., Kortmann, r., Koppenjan, J., Kwakkel, J., Ruijgh-van der Ploeg, T., Slinger, J., & Thissen, W. (2022). *Policy Analysis of Multi-Actor Systems* (Vol. 2). TU Delft Open. https://doi.org/ https://doi.org/10.5074/T.2022.004

Ersoy, S., & van der Gaag, S. (2023). *Studenten met migratieachtergrond opvallend vaak beschuldigd van fraude, minister wil systeem grondig nagaan*. NOS Nieuws. Retrieved June 25th 2023 from https://nos.nl/op3/artikel/2479700-studenten-met-migratieachtergrond-opvallend-vaak-beschuldigd-van-fraude-minister-wil-systeem-grondig-nagaan

European Commission. (2021). *Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain Union legislative Acts* (52021PC0206). https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A52021PC0206

European Commission Directorate-General for Communications Networks Content and Technology. (2019). *Ethics guidelines for trustworthy AI*. Publications Office,. https://doi.org/doi/10.2759/346720

European Union Agency for Fundamental Rights (FRA). (2022). *Bias in Algorithms: Artificial Intelligence and Discrimination*. Vienna Austria: European Union Agency for Fundamental Rights (FRA). https://fra.europa.eu/en/publication/2022/bias-algorithm

Faulconbridge, I., & Ryan, M. (2014). *Systems Engineering Practice*. Argos Press.

Gerards, J., Schäfer, M. T., Vankan, A., & Muis, I. (2021). *Impact Assessment Mensenrechten en Algoritmes*. https://www.rijksoverheid.nl/documenten/rapporten/2021/02/25/impact-assessment-mensenrechten-en-algoritmes

Govers, E., Hanse, D., van Beek, G., Mulder, J., & van de Wiel, J. (2021). *Een burger is geen dataset: Ombudsvisie op behoorlijk gebruik van data en algoritmen door de overheid* (2021/021). Nationale Ombudsman. https://www.nationaleombudsman.nl/system/files/bijlage/DEF%202.0%20Rapport%20%E2%80%93%20Een%20burger%20is%20geen%20dataset_0.pdf

Grimmelikhuijsen, S., & Meijer, A. (2020). Verantwoorde algoritmisering: zorgen waardengevoeligheid en transparantie voor meer vertrouwen in algoritmische besluitvorming? *Bestuurskunde, 2020*(4), 7-20. https://doi.org/10.5553/Bk/092733872020029004002

Grimmelikhuijsen, S., & Meijer, A. (2022). Legitimacy of Algorithmic Decision-Making: Six Threats and the Need for a Calibrated Institutional Response. *Persepectives on Public Management and Governance, 2022*(5), 232-242. https://doi.org/10.1093/ppmgov/gvac008

Groenlinks-PvdA. (2023). *Samen voor een hoopvolle toekomst: verkiezingsprogramma 2023*. https://groenlinkspvda.nl/verkiezingsprogramma/

Grondwet, (2023, 22 February). https://wetten.overheid.nl/BWBR0001840/2023-02-22/0#Hoofdstuk4

Hamer, J., Lemmens, A., & Kool, L. (2022). *Algoritmes Afwegen – Verkenning naar maatregelen ter bescherming van mensenrechten bij profilering in de uitvoering*. Rathenau Instituut. https://www.rathenau.nl/nl/digitalisering/algoritmes-afwegen

Harding, T., & Whitehead, D. (2013). Analysing data in qualitative research. In Z. Schneider, D. Whitehead, G. LoBiondo-Wood, & J. Haber (Eds.), *Nursing & Midwifery Research: Methods and Appraisal for Evidence-Based Practice* (pp. 141-160). Elsevier.

https://www.researchgate.net/publication/284372016_Analysing_data_in_qualitative_research

High-Level Expert Group on AI. (2020). *The Assessment List for Trustworthy Artificial Intelligence (ALTAI) for self assessment*. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

Hijink, M. (2022, May 7th, 2022). IND verzweeg een 'dikke error' met kennismigranten. *NRC Handelsblad,* E10-E11. https://www.nrc.nl/nieuws/2022/05/06/ind-verzweeg-een-dikke-error-met-kennismigranten-a4123661

Houtman, J. (2023). *'Sommige techbedrijven denken nog steeds regulering te kunnen vermijden'*. Het Financieele Dagblad Retrieved June 27th 2023 from https://fd.nl/tech-en-innovatie/1472012/sommige-techbedrijven-denken-nog-steeds-regulering-te-kunnen-vermijden

Houtzager, D., & Verbeek, S. (2022). *Gelijk recht doen: Deelrapport Sociale Zekerheid bij het parlementair onderzoek naar de mogelijkheden van de wetgever om discriminatie tegen te gaan*. https://www.eerstekamer.nl/overig/20220614/gelijk_recht_doen_deelrapport_3/f=/vltshvh4jzv3_opgemaakt.pdf

Kallio, H., Pietlä, A.-M., Johnson, M., & Kangasniemi, M. (2016). Systematic methodological review: developing a framework for qualitative semi-structured interview guide. *Journal of Advanced Nursing, 72*(12), 2954-2965. https://doi.org/10.1111/jan.13031

Kamensky, J. (2011). Managing the Complicated vs. the Complex. *The Business of Government*, (Fall/Winter 2011), 66-70. https://www.businessofgovernment.org/node/2299#overlay-context=magazine

Kuziemski, M., & Misuraca, G. (2020, 2020/07/01/). AI governance in the public sector: Three tales from the frontiers of automated decision-making in democratic settings. *Telecommunications Policy, 44*(6), 101976. https://doi.org/https://doi.org/10.1016/j.telpol.2020.101976

Leveson, N. (2012). *Engineering a Safer World: Systems Thinking Applied to Safety*. The MIT Press. https://doi.org/10.7551/mitpress/8179.001.0001

Leveson, N. (2020). Safety III: A Systems Approach to Safety and Resilience. http://sunnyday.mit.edu/safety-3.pdf

Leveson, N., & Thomas, J. (2018). *STPA Handbook*. https://psas.scripts.mit.edu/home/get_file.php?name=STPA_handbook.pdf

Liamputtong, P. (2008). Qualitative data analysis: conceptual and practical considerations. *Health Promotion Journal of Australia, 20*(2), 133-139. https://doi.org/10.1071/he09133

Manifestgroep. (2013). *MANIFESTgroep halfjaarraportage maart-september 2013*. https://vng.nl/files/vng/halfjaarrapportage_manifestgroep_nummer_3.pdf

Manifestgroep. (2022). *Uitvoeringsagenda 2022-2025*. Retrieved April 20th, 2024 from https://manifestgroep.nl/page/view/db959bfd-bc5c-4e1d-85b4-5a83e4a96582/uitvoeringsagenda-2022-2025

Ministerie van Binnenlandse Zaken en Koninkrijksrelaties. (2023a). *Handreiking Algoritmeregister: aan de slag met het Algoritmeregister*. https://www.digitaleoverheid.nl/document/handreiking-algoritmeregister/

Ministerie van Binnenlandse Zaken en Koninkrijksrelaties. (2023b). *Implementatiekader 'Verantwoorde inzet van algoritmen'*. https://www.rijksoverheid.nl/documenten/rapporten/2023/06/30/implementatiekader-verantwoorde-inzet-van-algoritmen

Ministerie van Sociale Zaken en Werkgelegenheid. (2021). *Toezicht op UWV en SVB - Uitgangspunten*. https://open.overheid.nl/repository/ronl-e2980582-4d6a-4633-aca6-2389e683bd74/1/pdf/toezicht-op-uwv-en-svb-uitgangspunten.pdf

Ministerie van Sociale Zaken en Werkgelegenheid. (2023a). *Jaarverslag en Slotwet Ministerie van Sociale Zaken en Werkgelegenheid 2022*. https://www.rijksoverheid.nl/binaries/rijksoverheid/documenten/jaarverslagen/2023/05/17/sociale-zaken-en-werkgelegenheid-2022/jaarverslag-szw-2022.pdf

Ministerie van Sociale Zaken en Werkgelegenheid. (2023b). *Stand van de uitvoering sociale zekerheid*. https://www.rijksoverheid.nl/documenten/publicaties/2023/06/21/stand-van-de-uitvoering-sociale-zekerheid-juni-2023

Nas, S., & Ouburg, S. (2022). *Inrichting algoritmetoezicht: Scenario's korte en lange termijn*. Privacy Company. https://www.tweedekamer.nl/downloads/document?id=2022D56237#:~:text=De%20Autoriteit%20Persoonsgegevens%20(hierna%3A%20AP,3%2C6%20miljoen%20vanaf%202026.

National Client Council. (2023a). *Jaarverslag 2022 Landelijke Cliëntenraad: Een stem geven aan burgers*. https://o8s9ala602u.b-cdn.net/1ezozem2ednj-230313-def-lcr-jaarverslag-2022.pdf

National Client Council. (2023b). *Visie op burgerparticipatie: Betrokkenheid van burgers met ervaringen*. https://o8s9ala602u.b-cdn.net/62t8xvacd8ho-230315-def-notitie-lcr-burgerparticipatie.pdf

National Client Council. (n.d.). *Landelijke Cliëntenraad (LCR): Bestaanszekerheid voor iedereen*. Retrieved February 1st, 2024 from https://www.landelijkeclientenraad.nl/de-lcr

Nationale Ombudsman. (2021a). *Overheid en algoritmen: wat u moet weten over de risico's*. https://www.nationaleombudsman.nl/professionals/nieuws/artikel/2021/overheid-en-algoritmen-wat-u-moet-weten-over-de-risicos

Nationale Ombudsman. (2021b). *Overheid, data en algoritmen*. https://magazinenationaleombudsman.h5mag.com/overheid_data_algoritmen/

Nationale Ombudsman. (2023). *Ombudsagenda 2023*. https://www.nationaleombudsman.nl/professionals/nieuws/nieuwsbericht/2023/ombudsagenda-2023

NSC. (2023). *Verkiezingsprogramma 2023. Tijd voor herstel: vertrouwen. Zekerheid. Perspectief.* https://partijnieuwsociaalcontract.nl/verkiezingsprogramma

Olsthoorn, P. (2016). *Big Data voor Fraudebestrijding*. https://www.wrr.nl/publicaties/working-papers/2016/04/28/big-data-voor-fraudebestrijding

Omzien naar elkaar, vooruitkijken naar de toekomst (Coalitieakkoord 2021-2025). (2021). (35 788 Bijlage). Tweede Kamer der Staten-Generaal. https://zoek.officielebekendmakingen.nl/kst-35788-77.html#ID-1009823-d36e70

Overheid.nl. (n.d.-a). *Het Algoritmeregister van de Nederlandse overheid*. Retrieved December 13th, 2023 from https://algoritmes.overheid.nl/nl

Overheid.nl. (n.d.-b). *Hoge Colleges van Staat*. Retrieved January 15, 2023 from https://www.overheid.nl/wie-vormen-de-overheid/hoge-colleges-van-staat

Parlementaire Enquêtecommissie Fraudebeleid en Dienstverlening. (2024). *Blind voor mens en recht*. https://www.tweedekamer.nl/nieuws/persberichten/rapport-parlementaire-enquetecommissie-fraudebestrijding-en-dienstverlening

Partij voor de Dieren. (2023). *Een wereld te herwinnen: Tweede Kamerverkiezingen 22 november 2023 verkiezingsprogramma*. https://www.partijvoordedieren.nl/al-onze-idealen

Partij voor de Vrijheid. (2023). *Nederlanders weer op 1: PVV verkiezingsprogramma 2023*. https://www.pvv.nl/verkiezingsprogramma.html

Peeters, R., & Widlak, A. (2023). Administrative exclusion in the infrastructure-level bureaucracy: The case of the Dutch daycare benefit scandal. *Public Administration Review, 83*(4), 863-877. https://doi.org/10.1111/puar.13615

Prins, C. (2021). Discriminerende Algoritmes. *Nederlands Juristenblad*, *96*(20), 1454. https://research.tilburguniversity.edu/en/publications/discriminerende-algoritmes

Prins, C., Sheikh, H., Schrijvers, E., de Jong, E., Steijns, M., & Bovens, M. (2021). *Opgave AI: De nieuwe systeemtechnologie*. https://www.wrr.nl/publicaties/rapporten/2021/11/11/opgave-ai-de-nieuwe-systeemtechnologie

Raub, M. (2018). Bots, Bias and Big Data: Artificial Intelligence, Algorithmic Bias and Disparate Impact Liability in Hiring Practices. *Arkansas Law Review, 7*(12), 529-570. https://scholarworks.uark.edu/alr/vol71/iss2/7

Rechtbank Den Haag. (2020). *SyRI-wetgeving in strijd met het Europees Verdrag voor de Rechten voor de Mens*. Retrieved 25th June 2023 from https://www.rechtspraak.nl/Organisatie-en-contact/Organisatie/Rechtbanken/Rechtbank-Den-Haag/Nieuws/Paginas/SyRI-wetgeving-in-strijd-met-het-Europees-Verdrag-voor-de-Rechten-voor-de-Mens.aspx

Rengers, M., Houtekamer, C., & Maleeyakul, N. (2023). *'Pas op met deze visumaanvraag', waarschuwt het algoritme dat discriminatie in de hand werkt. Het ministerie negeert kritiek*. NRC Handelsblad. Retrieved May 6th, 2023 from https://www.nrc.nl/nieuws/2023/04/23/beslisambtenarenblijven-profileren-met-risicoscores-a4162837

Rijksinspecties. (2021). *Algoritmen en Artificiële Intelligentie: hoe houd je daar toezicht op?* https://www.rijksinspecties.nl/actueel/nieuws/2021/02/03/algoritmen-en-artificiele-intelligentie-hoe-houd-je-daar-toezicht-op

Rijksinspecties. (n.d.). *Over de rijksinspecties*. Retrieved February 1st, 2024 from https://www.rijksinspecties.nl/over-de-inspectieraad/over-de-rijksinspecties

Rijksoverheid (2022). *Werkagenda Waardengedreven Digitaliseren*. https://www.tweedekamer.nl/kamerstukken/brieven_regering/detail?id=2022D45419&did=2022D45419

Sargut, G., & McGrath, R. (2011). Learning to Live with Complexity. *Harvard Business Review*, (September 2011). https://hbr.org/2011/09/learning-to-live-with-complexity

Schelfhout, D., Geurts, S., de Craen, E., van Veldhuisen, A., Engelen, M., & Drouen, T. (2022). *Grip op gegevensuitwisseling: Eindrapport onderzoek naar de governance van de Stichting Inlichtingenbureau en BKWI*. https://www.tweedekamer.nl/downloads/document?id=2023D00390

Schwartz, R., Vassilev, A., Greene, K., Perine, L., Burt, A., & Hall, P. (2022). *Towards a Standard for Identifying and Managing Bias in Artificial Intelligence*. National Institute of Standards and Technology (NIST). https://doi.org/10.6028/NIST.SP.1270

Sociale Verzekeringsbank. (2020). *IV-Strategie 2021-2025*. https://vng.nl/sites/default/files/2020-12/iv-strategie_svb_2021_-_2025.pdf

Sociale Verzekeringsbank. (2023). *Meerjarenplan Handhaving 2023-2026: Handhaven vanuit vertrouwen*. https://www.rijksoverheid.nl/documenten/kamerstukken/2023/01/11/bijlage-4-meerjarenplan-handhaving-svb-2023-2026

Sociale Verzekeringsbank. (n.d.-a). *Cliëntenraad SVB*. Retrieved February 1st, 2024 from https://www.svb.nl/nl/over-de-svb/wie-zijn-we/clientenraad-svb

Sociale Verzekeringsbank. (n.d.-b). *De SVB voor Sonja, Monique, Mehmet, Jakko, Jack en Guusje: Meerjarenkoers 2021-2025*. https://www.svb.nl/nl/media/SVB-meerjarentraject-2021-2025-web.pdf

Sociale Verzekeringsbank. (n.d.-c). *Hoe gaan we om met algoritmes*. Retrieved February 1st, 2024 from https://www.svb.nl/nl/over-de-svb/hoe-werken-we/hoe-gaan-we-om-met-algoritmes

Sociale Verzekeringsbank. (n.d.-d). *The story of the SVB*. Retrieved December 14th, 2023 from https://www.svb.nl/en/about-the-svb/who-we-are/the-story-of-the-svb

Socialistische Partij. (2023). *Nu de mensen: programma van de socialistische partij voor de tweede kamerverkiezingen van 22 november 2023*. https://www.sp.nl/verkiezingsprogramma2023

Stichting Kafkabrigade. (2023). *Naar een handelingsperspectief - handvatten bij onderzoek stelselkwaliteiten*. https://staatvandeuitvoering.nl/app/uploads/2022/06/2023019-handelingsperspectief_stelselkwaliteiten.pdf

Swierstra, T., & Vermaas, P. (2022). 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* (pp. 239-268). Uitgeverij Boom.

Tijdelijke Commissie Uitvoeringsorganisaties. (2021). *Klem tussen balie en beleid*. Tweede Kamer der Staten-Generaal. https://www.tweedekamer.nl/kamerleden-en-commissies/commissies/tijdelijke-commissie-uitvoeringsorganisaties/eindrapport

Tweede Kamer der Staten-Generaal. (2023). *Taken en rechten*. https://www.tweedekamer.nl/sites/default/files/field_uploads/factsheet%20Taken%20en%20rechten%20(2013)_tcm181-180347.pdf

Tweede Kamer der Staten-Generaal. (n.d.). *Fracties*. Retrieved March 1st, 2024 from https://www.tweedekamer.nl/kamerleden_en_commissies/fracties

Uitvoeringsinstituut Werknemersverzekeringen. (2021a). *Beleidsdocument model risico management*. https://www.uwv.nl/overuwv/Images/bijlage-4-beslissing-op-bezwaar-op-wob-verzoek-software-en-algoritmes.pdf

Uitvoeringsinstituut Werknemersverzekeringen. (2021b). *Model ontwikkeling - Model documentatie Verblijf Buitenland*. https://www.uwv.nl/overuwv/Images/bijlagen-bij-deelbesluit-2-deel-2.pdf

Uitvoeringsinstituut Werknemersverzekeringen. (2021c). *UWV Kompas Data Ethiek*. https://www.uwv.nl/imagesdxa/kompas-data-ethiek_tcm94-438618.pdf

Uitvoeringsinstituut Werknemersverzekeringen. (2022). *Besluiten- en actiepuntenlijst Vergadering Directie WERKbedrijf*. https://www.uwv.nl/overuwv/Images/bijlagen-bij-deelbesluit-2-deel-1.pdf

Uitvoeringsinstituut Werknemersverzekeringen. (2023a). *UWV Informatieplan 2023-2027: Werken aan verbeteren dienstverlening en ICT-fundament*. https://open.overheid.nl/repository/ronl-5aef67465502fda97691b20ad30eca8be05362c2/1/pdf/bijlage-5-uwv-informatieplan-2023-2027.pdf

Uitvoeringsinstituut Werknemersverzekeringen. (2023b). *UWV jaarverslag 2022, deel 1*. https://jaarverslag.uwv.nl/FbContent.ashx/pub_1000/downloads/v230420132204/uwv-jaarverslag-2022-deel1.pdf

uitvoeringsinstituut Werknemersverzekeringen. (n.d.). *Algoritmeregister UWV*. Retrieved February 1st, 2024 from https://www.uwv.nl/nl/over-uwv/organisatie/algoritmeregister-uwv

Uitvoeringswet Algemene verordening gegevensbescherming, (2021, 1 July). https://wetten.overheid.nl/BWBR0040940/2021-07-01#Hoofdstuk2

van Atteveldt, W., Roozendaal, W., Ruigrok, N., de Vries, K., van der Velden, M., Busuioc, M., & Guldemond, P. (2024). *Tussen Ambitie en Uitvoering. Een contextanalyse van de dynamiek tussen media, politiek en beleid bij de totstandkoming en uitvoering van dertig jaar sociale zekerheid*. https://www.tweedekamer.nl/kamerstukken/detail?id=2024D06759&did=2024D06759

van der Sloot, B., Keymolen, E., Noorman, M., College voor de Rechten van de Mens, Weerts, H., Wagensveld, Y., & Visser, B. (2021). *non-discriminatie by design*. https://www.rijksoverheid.nl/documenten/rapporten/2021/06/10/handreiking-non-discriminatie-by-design

van Dooijeweert, A. (2021). *Advies inzake het wetsvoorstel Wet gegevensverweking door samenwerkingsverbanden*. College voor de Rechten van de Mens. https://publicaties.mensenrechten.nl/publicatie/60d5838b98d7821c6468361a

van Huffelen, A. C. (2022a). *Inrichtingsnota Algoritmetoezichthouder* (kst-26643-953). https://www.tweedekamer.nl/kamerstukken/brieven_regering/detail?id=2022Z26092&did=2022D56234

van Huffelen, A. C. (2022b). *Kamerbrief werkagenda waardengedreven digitaliseren* (kst-26643-940). https://www.tweedekamer.nl/kamerstukken/brieven_regering/detail?id=2022Z21101&did=2022D45419

van Huffelen, A. C. (2023a). *Opschaling Algoritmetoezichthouder (Directie Coördinatie Algoritmes bij de AP)* (2023-0000738933). https://www.rijksoverheid.nl/documenten/kamerstukken/2023/12/21/kamerbrief-over-opschaling-algoritmetoezichthouder-directie-coordinatie-algoritmes-bij-de-ap

van Huffelen, A. C. (2023b). *Toezicht in het digitale domein (o.a. op AI en algoritmes)* (2023-0000235995). https://www.rijksoverheid.nl/documenten/kamerstukken/2023/05/24/kamerbrief-over-toezicht-in-het-digitale-domein-o-a-op-ai-en-algoritmes

Versmissen, K., & Soerjadi, A. (2022). *Data-ethiek in uitvoering*. Expertisecentrum Data-Ethiek. https://staatvandeuitvoering.nl/onderzoek/data-ethiek-in-uitvoering/

Vié, J. (2023). *'Studenten met migratieachtergrond vaker beschuldigd van fraude door DUO'*. NRC Handelsblad. Retrieved 25th June 2023 from https://www.nrc.nl/nieuws/2023/06/21/studenten-met-migratieachtergrond-vaker-beschuldigd-van-fraude-door-duo-a4167720

VVD. (2023). *Ruimte geven. Grenzen stellen. Keuzes voor een optimistische toekomst. Verkiezingsprogramma VVD 2023*. https://www.vvd.nl/nieuws/definitief-verkiezingsprogramma/

Wet College voor de rechten van de mens, (2020, 1 January). https://wetten.overheid.nl/BWBR0030733/2020-01-01/0#Hoofdstuk6

Wet structuur uitvoeringsorganisatie werk en inkomen, (January 1, 2024). https://wetten.overheid.nl/BWBR0013060/2024-01-01

Widlak, A. (2022). *Overzicht & Ondergrens: digitale overheid in kort bestek*. Stichting Kafkabrigade. https://staatvandeuitvoering.nl/onderzoek/overzicht-ondergrens-digitale-overheid-in-kort-bestek/

Wieringa, M. (2020). What to account for when accounting for algorithms. Conference on Fairness, Accountability, and transparency (FAT '20), Barcelona, Spain. https://dspace.library.uu.nl/bitstream/handle/1874/414522/3351095.3372833.pdf?sequence&

Wieringa, M. (2023). "Hey SyRI, tell me about algorithmic accountability": Lessons from a landmark case. *Data & Policy, 5*, e2, Article e2. https://doi.org/10.1017/dap.2022.39

Zenger, R. (2017). *Zeg 'nee' tegen 'computer says no'*. Bits of Freedom. https://www.bitsoffreedom.nl/2017/02/08/zeg-nee-tegen-computer-says-no/

Zouridis, S., Eck, M. v., & Bovens, M. (2019). Automated Discretion. In P. Hupe & T. Evens (Eds.), *Forthcoming, Palgrave Handbook on discretion: The Quest for Controlled Freedom*. https://dx.doi.org/10.2139/ssrn.3453068

Appendix A: Actor Analysis

This appendix includes additional information for the actor analysis.

A.1 Analysis of themes in party programmes

Table A 1 displays the ten largest political parties of the 2024 Dutch national election, together with themes found in their party programmes that are relevant for the actor analysis.

Table A 1 Relevant themes in party programmes of the biggest parties in the 2024 elections

Political party	Relevant themes
Partij voor de Vrijheid (2023)	Livelihood, tackle abuse of social security policies.
Groenlinks-PvdA (2023)	Livelihood, lack of trust in government, distrustful government, take citizen viewpoint into account more explicitly, approachable government (non-digital), complexity of rules, discriminating systems, AP as algorithm watchdog, digital inclusion, minister for digitalisation.
VVD (2023)	Simplification of rules and regulations, abolish benefits system, human dimension, honest use of algorithms (incl. transparency).
NSC (2023)	Livelihoods, distrustful government, lack of trust in government, non-digital contact, funds for AP, more transparency when using algorithms (register), digital skills.
D66 (2023)	Complexity of rules and regulations, mandatory use of algorithm register and human-rights tests, digital inclusion, more funds for AP, minister for digitalisation, abolish benefits system.
BBB (2023)	Transparency of digital means, minister for digitalisation.
CDA (2023)	Livelihoods, distrustful government, complexity of rules and regulations, right to knowing data and algorithms government uses.
Socialistische Partij (2023)	Distrustful government, digitalisation and dehumanisation, abolish benefits system, forbid fraud-detection AI & algorithms if human rights are impacted.
DENK (2023)	Risk models and discrimination, more funds for AP, mandatory algorithm register, human dimension, distrustful government, simplification of system.
Partij voor de Dieren (2023)	Distrustful government, digital rights and threats of digitalisation, non-digital contact with government.

A.2 Analysis of actor interests

Table A 2 displays the different included actors and their interests, objectives, as well as views on the problem and actions they take within the problem situation.

Table A 2 Analysis of actor interests

Actors	Interests	Objectives	Existing Situation, Gap	Causes	Current Actions/Used Means
SVB	Execution of social insurance schemes: informing, judging, payment, enforcement of laws.	Working in a trustworthy and innovative way with eye for the human dimension (Sociale Verzekeringsbank, n.d.-b). Effective use of capacity by use of digital means (Sociale Verzekeringsbank, 2023).	Use of algorithms can cause risks (exclusion, errors, discrimination & profiling, lack of transparency)	Level of (digital) literacy in citizens, standardised approach, errors in data (exchange), complexity of systems and overarching rules.	• Legal and ethical tests • Development and certification of algorithms with external experts • Normative framework • SVB algorithm register and National algorithm register • Intention workshops (Garages de bedoeling) • Citizen panels before use of algorithm • Consultation of Advisory Board, IT Committee and ethics centre • Human-in-the-loop
UWV	Execution of employee insurance schemes.	Using IT to provide better public service with fewer employees.	Use of algorithms can impact citizens and organisation (impact on citizens, reputation damage, privacy breaches, bad decision-making, financial impacts).	Model Risks (errors) leading to inaccurate(ly interpretated) results influencing decision-making in the organisation.	• Human-in-the-loop • Compass data-ethics; tool for responsible use of data and algorithms • UWV Algorithm register and National algorithm register • Consultation of commission data ethics when dilemmas arise • Data academy: education on relevant themes • Cooperation between SVB, UWV, SZW regarding information provision • Model Risk Management; policy for risk scans
Ministry of Social Affairs	Create policy to support citizens when it comes	Responsible for execution of social security laws by	Illegitimate use of social security policies,	Complexity of laws and regulations, insufficient law	• Cooperation with SVB, UWV, LCR including reporting such as

	to livelihood and labour participation (Ministerie van Sociale Zaken en Werkgelegenheid, 2023a, 2023b).	executing agencies (UWV, SVB).	policies not accessible to all citizens, citizens stuck in bad situations (e.g. unsolved issues, debts) (Ministerie van Sociale Zaken en Werkgelegenheid, 2023a)	enforcement instruments, citizen not central in policy execution/human dimension not applied (Ministerie van Sociale Zaken en Werkgelegenheid, 2023b).	bottleneck letters and yearly reports • Programmes to improve execution (e.g. 'Werk aan uitvoering') • Shaping law and policy with which UWV and SVB operate • Organisational oversight over UWV/SVB
Ministry of Internal Affairs	Interests: securing public values in the digital domain, creating a level playing field in the economy, acts as 'client' of DCA.	Inclusivity in digital era, trust in digital world, control on digital world, digital government that works in value-driven and open manner for all citizens, utilize algorithmic systems for public good (Rijksoverheid 2022; van Huffelen, 2022b).	Public values and basic rights are not protected well enough, use of algorithms is not in control well enough. Problematic cases still arise in various sectors. (Rijksoverheid 2022; van Huffelen, 2022b)	Oversight over algorithm use isn't structural (enough) in nature (Rijksoverheid 2022).	Various actions under agenda value-driven digitalization: • Appointing DCA as NAC • National algorithm register • Coordination of digital laws and policy across government (within which DCA also operates) • IAMA • Frameworks (Implementation framework, Algorithm framework.
Dutch DPA (part of AP)	Supervision over compliance with GDPR (general purpose)	Protect personal data of citizens.	Personal data not always optimally protected in digitalising society (Autoriteit Persoonsgegevens, 2023a).	Various causes, lack of knowledge, lack of willingness, increasing digitalisation.	• Preventative oversight (guidance through guidelines, conversations, reports, supporting professionals), • Repressive oversight (sanctioning, fines) • Advising on laws and regulations by government • Processing and investigating reports of data leaks • Processing and investigating citizen complaints.
Directorate Coordination Algorithms (DCA) (part of AP)	National Algorithm Coordinator in the Netherlands, coordination of supervision over	Protecting public values and basic human rights when algorithmic systems are used (e.g discrimination, arbitrariness, deception,	Use of algorithmic systems can cause risks to individuals and groups, and impact public values and basic	Various reasons, such as lack of structural and institutional oversight, incongruent development of oversight and risk control, lack of	• Risk signalling and explanation shared norms and values through publication (e.g. risk reports • Cooperation with supervisors

	algorithmic systems.	lack of transparency, explainability, freedom of speech, equality of opportunity) (Autoriteit Persoonsgegevens, 2023c, n.d.) Focus on transparency, auditing, governance and discrimination (Autoriteit Persoonsgegevens, 2024).	rights.	transparency, lack of binding rules, frameworks and guidelines, risk control initiatives are just beginning to take place (Autoriteit Persoonsgegevens, 2023b, 2023c).	and supervised organisations (e.g. panels, conversations, surveys)
Netherlands Institute of Human Rights (CRM)	Protecting human rights, increase attention for and compliance with these rights, especially right of equal treatment and non-discrimination.	Protecting human rights when it comes to algorithm use, increase awareness for human rights.	Use of algorithms and (semi-) automated decision-making still impacts human rights (incl. exclusion, discrimination, freedom of speech, privacy).	Wider organisational flaws (lack of safeguards to prevent discrimination, lack of attention for human rights) and specific flaws within algorithmic systems (skewed datasets).	• Advise government(al organisations) • Investigations into human right violations, resulting in ending in judgments • Essays, articles etc. • Cooperation with other parties
National Ombudsman	Provide help in situations where something between government and citizens goes wrong, focus on conduct of government.	Clear, open, and solution-oriented government conduct regarding algorithm use. Helping citizens who are in conflict with government.	Unwanted effects of algorithm use by government (e.g. benefits scandal) (Govers et al., 2021). Examples are discrimination, privacy violations, lack of transparency, lack of service fitting to needs of individuals (non-standardized aid) (Nationale Ombudsman, 2021a).	Citizen perspective is not always explicitly named or taken into account in policy and government actions (Govers et al., 2021). Citizens aren't sufficiently involved in algorithm development (Nationale Ombudsman, 2021b).	• Magazines, visions, news articles • Encourage organisations to develop legal and ethical frameworks and apply the human dimension • Investigate citizen complaints and publish judgments
Court of Audit (ARK)	Ensuring public accountability over spendings of public funds, efficient and	Efficient and purposeful policy around algorithm use by governmental organisations.	(Uncareful) use of algorithms by governments brings about risks for citizens	Lack of overall control over threats (Algemene Rekenkamer, 2021a).	• Investigations into government spending and policy • Yearly reports to Chamber of Representatives

	purposeful government policy and spendings.		and society (e.g. lack of explainability, discrimination, lack of transparency) (Algemene Rekenkamer, 2021a, 2022).	Several underlying causes for threats, such as complexity of technology, skewed data, lack of lifecycle management, and so on (Algemene Rekenkamer, 2021a, 2022).	• Algorithm assessment framework, also used in reports and investigations
Amnesty International (Netherlands division)	Ensuring internationally recognized human rights are respected (Amnesty International, 2023c).	Ensuring rights of the International Covenant on Civil and Political Rights when technology is used (Amnesty International, 2023c).	Human rights of citizens are impacted by the use of technology by government(al agencies) (Amnesty International, 2021, 2023b, 2023c).	Government use of technology doesn't sufficiently keep human rights into account (Amnesty International, 2023c). Government doesn't do enough to protect citizens' rights (Amnesty International, 2023b).	• Lobbying in EU and NL • Creating awareness • Investigations • Cooperation with other NGOs • Training government employees about risk profiling • Dialogues with government agencies
Bits of Freedom	Ensuring an open and just information society (Bits of Freedom, n.d.-a).	Specific focus on freedom of communication and privacy (Bits of Freedom, n.d.-b).	Lack of transparency or explainability when government uses algorithms (Zenger, 2017).	Complexity of used algorithms (Zenger, 2017).	• Campaigning • Legal actions • Advocacy
House of Representatives (Tweede Kamer)	Representing citizens, shaping law and scrutinizing government actions (specific interests vary between parties).	Objectives vary between parties, but common themes include transparency of algorithm use, distrustful government, human dimension and livelihoods, complexity of social security system.	Various; includes lack of transparency, impacts on human rights, eroded relationship between citizens and government, lack of livelihood security, abuse of social security schemes.	Complexity of system, lack of digital skills, distrustful government, lack of oversight over algorithms, human dimension not central, lacking oversight over abuse of social security schemes.	• Budgeting rights • Right to amend laws • Right to inquire (parliamentary inquiries) • Right to propose motions • Right of interpellation • Right to propose laws • Debates, discussions • Parliamentary committees
National Client Council (LCR)	Representing citizens that receive government aids.	Citizen participation, flaws in execution, accessibility of system (National Client Council, 2023a, 2023b, n.d.).	Citizens do not trust government, get into trouble in the system, system isn't accessible enough (National Client Council, 2023a).	Strict and complicated laws and regulations, citizen perspective not sufficiently taken into account (National Client Council, 2023a).	• Consultation of UWV, SVB, SZW • Citizen tests • Citizen journeys • Discussion with clients • Member of several commissions working to help clients and investigate improvements

					• Contribute to political debate • Signalling bottlenecks • Inform clients • Public campaigns • Request information (UWV, SVB) • Advise UWV, SVB (solicited and unsolicited) • Propose plans and adjustments to SVB, UWV, SZW
European Union (EU)	Supranational lawmaker	Various, includes protecting human rights and internal free market.	AS impact human rights and economy, are difficult to control.	Lack of (unified) regulation within EU.	• Supranational law (GDPR, AI Act)

Appendix B: System-Theoretical Process Analysis

This appendix includes additional information used for STPA.

B.1 Selection of AS

Table B 1 displays the different AS found in the algorithm registers of UWV and SVB, whether they were included or not, and the rationale for this.

Table B 1 Selection of AS for STPA

Organisation	Name of algorithmic system	Purpose	Included	Rationale
UWV	Risicoscan verblijf buiten Nederland	Signalling whether those receiving unemployment benefits are staying abroad. (Out of use)	Yes	Citizens who stayed abroad and didn't adhere to rules might not be entitled to benefits. The risk scan therefore influenced the benefits scheme.
	Claim Beoordelings- en Borgingssysteem (CBBS)	System helps determine disability percentage.	Yes	Disability percentage influences whether and how much benefits one receives.
	Klantapplicatie WW	Helps citizen file unemployment benefit request, and afterwards automatically decides whether citizen is entitled to benefits.	Yes	Algorithmic system automatically decides and therefore influences benefits payment.
	Digitale vragenlijst ziekmelding	Adjusts questionnaire regarding sickness benefits, only showing relevant questions.	Yes	Algorithm influences the questions that are asked, and the answers to questions influences entitlement to benefits. The algorithm can therefore indirectly influence benefits payment.
	Werkverkenner	Helps decide if citizen needs extra guidance in finding work.	No	System only supports guidance of citizens, but has no influence on benefits one receives.
	Sollicitatiescan WW	Helps decide if citizen needs extra guidance in applying for jobs.	No	System only supports guidance of citizens, but has no influence on benefits one receives.
	Maatwerkscan	Judges whether citizens run risk of falling into social assistance scheme after receiving unemployment benefits.	No	System only supports guidance of citizens, but has no influence on benefits one receives.
	Risicoscan verwijtbare werkloosheid	Judges whether citizen became unemployed outside of their own fault.	Yes	When citizen is not at fault, they are not entitled to benefits. The risk scan can therefore influence

				entitlement to benefits.
	Vaststellingmodel regres	Judges whether a third party might be (financially) responsible for citizen becoming unable to work (disability).	No	Algorithmic system only influences who might pay for disability benefits, not whether a citizen is entitled to them or not.
SVB	Algoritmes AOW	Judges whether someone is entitled to basic state pension and automatically handles payment.	Yes	Algorithm influences payment of basic state pension.
	Algoritmes kinderbijslag	Judges whether family is entitled to child benefits and automatically handles payment.	Yes	Algorithm influences payment of child benefits scheme.
	Algoritmes AIO-aanvulling	Judges whether and how much additional income for elderly one receives, and if this is a benefit or a loan.	Yes	Algorithm influences payment of additional income for elderly.
	Algoritmes sancties	Decision tool that helps employees judge whether punishment is in order (e.g. when too much benefits have been received).	Yes	When punishment is in order, this can influence the finances of citizens and the payment of benefits.
	Algoritmes vrijwillige verzekering AOW en Anw	Calculates how much premium needs to be paid when citizen moves abroad and wants to stay insured for state and relative pensions (AOW, AnW)	No	Algorithm influences financial situation of citizen, however this is optional. Too little information to contribute to analysis.
	Algoritmes Preventie en Handhaving	SWAN-model is an AI-model that is not yet in use. It predicts whether those who receive benefits live together even though they state that they live alone.	Yes	When an investigation finds a citizen lives together, this influences their entitlement to benefits.

B.2 hazardous control actions

This appendix lists the different control actions that can lead to hazards.

B.2.1 Hazard 1 : flawed logic

Table B 2 Hazardous control actions for hazard 1 (flawed logic)

Control action	Type	Link to hazard	SCS
Monitoring of model outputs and use	Providing causes hazard	Monitoring based on KPIs can limit insight into factors at play	UWV
	Not providing causes hazard	No monitoring cuts out all insight into workings of AS	
Development and maintenance by developers	Providing causes hazard	Development based on limited insight in relevant factors can translate flawed logic into AS.	UWV
	Not providing causes hazard	No maintenance leaves out changes needed to reflect context changes.	
	Applied too late	Adapting AS to reflect changes in laws, policies, or other relevant context can cause logic to be outdated.	
Requesting advice of UWV client council	Not providing causes hazard	Not seeking advice can limit ability to keep into account relevant factors.	UWV
	Stopped too early	Advice requested only during development and not during use can limit insight into effects while in use.	
Requesting advice of external organisations	Providing causes hazard	Focus on legal and ethical aspects can limit insight into impacts on citizens.	UWV/SVB
	Not providing causes hazard	No external advice can cause biases and flawed logic to go unnoticed.	
Requesting advice of ethical committee	Providing causes hazard	Focus on ethical aspects can limit insight into impacts on citizens.	UWV
	Not providing causes	No advice can leave out ethical	

	hazard	aspects or external points of view.	
Provision of personalised questionnaire (digital sickness checklist UWV) or personalised request form (child benefits SVB) to citizen	Providing causes hazard	Providing personalised questionnaires or request forms based on flawed logic can lead to non-fitting forms and questionnaires, impacting benefits allocation.	UWV
Risk estimation by risk scan	Providing causes hazard	Risk estimation based on flawed logic can cause citizens to be flagged unjustly and/or disproportionally.	UWV/SVB
Decision made by executing employee	Providing causes hazard Applied too late	Decision based on flawed information or flawed logic can cause flawed allocation of benefits Late decisions due to mismatched logic needing employee judgment can lead to citizens being helped inefficiently	UWV/SVB
Investigation on citizen by executing employee	Providing causes hazard	Investigations based on flawed logic can cause citizens to be investigated unjustly and/or disproportionally.	UWV/SVB
Measure taken on citizen	Providing causes hazard	Measures based on flawed logic can be unjust or disproportionate	UWV/SVB
Instructions and internal policy provided by higher-up management	Providing causes hazard Not providing causes hazard	Instructions can influence development and use, including consultation with other parties, attention for citizen impacts and discretionary space. Limited or no instructions can cause errors in different phases of AS lifecycle, lack of carefulness and arbitrariness.	UWV/SVB
Requesting advice from citizen panel	Not providing causes hazard Stopped too soon	Advice only asked when impact is expected, leaving out unexpected impacts. Advice is only asked during development, not during use.	SVB
Law and policy creation by ministries and chamber of representatives	Providing causes hazard Not providing causes hazard	Policy can influence internal policy of organisations, use of AS, and supervision thereof. Limited or no law and policy can cause hazards due to lacking internal policy, errors in use, lacking supervision.	General
Requesting advice from	Not providing causes	No or limited requests for advice	General

LCR	hazard	from LCR can limit insight into important factors.	
Reporting by press	Providing causes hazard Not providing causes hazard	Reporting can cause focus on fraud prevention and detection, and cause distrust in citizens. Reporting by the press can uncover scandals, inform about flawed practices within organisations, and about influences on citizens.	General

B.2.2 Hazard 2: Flawed information used in AS

Table B 3 Hazardous control actions for hazard 2 (flawed information)

Control action	Type	Link to hazard	SCS
Risk estimation by risk scan	Providing causes hazard Not providing causes hazard	Risk estimation based on flawed information can cause citizens to be flagged unjustly and/or disproportionally. Not doing a risk scan due to flawed information estimation limit capacity to enforce rules.	UWV/SVB
Decision made by executing employee	Providing causes hazard	Decision based on flawed information or flawed logic can cause flawed allocation of benefits	UWV/SVB
Investigation on citizen by executing employee	Providing causes hazard Not providing causes hazard	Investigations based on flawed information or flawed logic can cause citizens to be investigated unjustly and/or disproportionally. Not doing an investigation because flawed information is present can mean rules are not enforced and citizens make unjust use of benefits.	UWV/SVB
Measure taken on citizen	Providing causes hazard	Measures based on flawed information or flawed logic can be unjust or disproportionate	UWV/SVB
Requesting information from citizen	Providing causes hazard	Scoping of information request only allows citizen to provide information requested and can leave out important information and context, especially when requests are standardised and discretionary space is limited. Not asking citizen for information	UWV/SVB

	Not providing causes hazard	can leave out necessary information or ability to correct information.	
Requesting information from citizen through personalised application form	Providing causes hazard	Scoping of information request through automated means only allows citizen to provide information requested and can leave out important information and context.	UWV/SVB
Using pre-filled forms	Providing causes hazard	Erroneous information can be overlooked (overreliance by citizen).	UWV/SVB
Information request/automatic data transfer through SUWInet	Providing causes hazard	Erroneous info can permeate through system, limited authority over information, different use cases (and definitions), differences in currentness of data sources.	UWV/SVB/general
Instructions and internal policy provided by higher-up management	Providing causes hazard Not providing causes hazard	Internal policy can impact scrutiny, data sharing, standardisation of data requests and discretionary space. Limited or no internal policy can cause lack of carefulness or arbitrariness.	UWV/SVB
Law and policy creation by ministries and chamber of representatives	Providing causes hazard Not providing causes hazard	Policy can influence internal policy of organisations, use of AS, and supervision thereof. No or limited policy can cause hazards due to lacking internal policy, use of data, and supervision thereof.	General

B.2.3 Hazard 3: Citizen has flawed understanding

Table B 4 Hazardous control actions for hazard 3 (flawed understanding)

Control action	Type	Link to hazard	SCS
Information about decisions or investigations provided to citizens.	Providing causes hazard Not providing causes hazard Too late	Limited information about use of AS provided can cause citizen to lack information. No information will lead to citizen having little to no understanding about what has happened. Providing information too late can impact actionability of information and/or cause citizen	UWV/SVB/General

		to be in dire situation.	
Requesting advice from client council.	Not providing causes hazard	Client council can judge on understandability of AS.	UWV
Instructions and internal policy provided by higher-up management	Providing causes hazard Not providing causes hazard	Instructions can influence information given to citizens, complexity of AS, and advice requests. No internal policy can cause flaws in processes and minimal information to arise, as well as arbitrariness.	UWV/SVB
Reporting by press	Providing causes hazard Not providing causes hazard	Reporting with focus on scandals can instil fear within citizens. Press provides extra oversight and can uncover scandals.	General
Information requests through SUWInet	Providing causes hazard	Data coupling increases system complexity and thereby understandability.	UWV/SVB/General
Law and policy creation by chamber of representatives and ministries	Providing causes hazard Not providing causes hazard	Stacking of laws and policies increases complexity and decreases understandability. Laws and policies on transparency can influence information available to citizen (e.g. register). Lacking law on transparency can cause little information to be available (e.g. non-mandatory algorithm register).	General
Information provided by algorithm register	Providing causes hazard Not providing causes hazard	Register is incomplete and holds little (actionable) information. Register does give some centralised information about use of AS that would otherwise be missing.	General

Appendix C: Interview setup

This appendix shows the interview protocol, as well as the information used in preparation for the interviews.

C.1 Interview protocol

Below are the general four themes used in interviews and the types of questions that were asked within these themes. This served as a guide for conducting interviews. These questions were changed according to the actor that was interviewed, the information used for this can be found in Table C 1.

1. What does your organisation do and what is your role?
    a. How do you gather information needed (means, active, passive)?
    b. What other actors are you in contact with?
    c. What communication streams are there with these other actors?
    d. What communication streams are there within your organisation?
    e. What are (recent) artefacts produced by your organisation?
2. What do you see in your role?
    a. What goes well in the governance of AS (at executing agencies)?
    b. What issues are still present in governance of AS (at executing agencies) and why?
    c. How do you see recent developments (law, register, DCA, etc.)?
    d. What does existing law and regulation provide and what not?
    e. What are gaps in actions and law?
    f. What were responses to publications and actions by your organisation?
3. What needs to happen within organisations?
    a. Relating back to issues (2), what needs to happen within organisations?
    b. What recommendations has your organisation made?
    c. What does your organisation need to better execute their role (e.g. means, information)?
    d. What do executing agencies need to do to prevent harms, and what do they need in order to do this?
4. What needs to happen within the wider supervisory landscape?
    a. Relating back to issues (2), what needs to happen in the wider landscape?
    b. Relating back to recent actions and gaps, what is needed to close these

C.2 Information used for interviews

Table C 1 Specification of themes and sources used for interviews

Respondent	Specific themes (from AA)	Sources
Researcher National Ombudsman	• Ombudsvisie • Publications on algorithms and citizens • Participation of and role of citizen	(Govers et al., 2021; Nationale Ombudsman, 2021a, 2021b)
Policy Advisor Amnesty Netherlands	• Focus on risk profiles and discrimination • Reporting on risk profiles • Reporting on supervision over algorithms and risk profiles and scandals	(Amnesty International, 2021, 2023a, 2023b, 2023c)
Leadership role Enforcement Division UWV	• Model Risk Management and roles therein in practice • 70/30-rule • Supervision by Social Affairs (SZW) • Transparency: register, gaming the system • Checking assumptions	(Uitvoeringsinstituut Werknemersverzekeringen, 2021a, 2021b, 2022, 2023a, 2023b, n.d.)
Researcher Court of Audit	• Algorithm Assessment framework • Reporting on topic of algorithms • Role of ARK in supervisory landscape	(Algemene Rekenkamer, 2021a, 2021b, 2022, n.d.-b)
Consultant National IT Guild (RIG)	N/A	N/A
Policy Advisor CvdRM	• Assessment framework • Paper on transparency	(College voor de Rechten van de Mens, 2021, 2023)
Employee National Client Council	• Connection with citizens • Connection with UWV/SVB and client councils	-
Policy Advisor AI and Algorithms BZK	• Value-driven digitalisation agenda • Algorithm framework • Algorithm register, transparency	(Digitale Overheid, 2023; Overheid.nl, n.d.-a; Rijksoverheid 2022)

Appendix D: Interview reports

This appendix includes reports of the interviews that were elicited. Interviews have been summarised slightly, still leaving important context in the summary. Functions of respondents have been generalised in order to decrease possibility of traceability.

D.1 Researcher at National Ombudsman (I1)

What does NO do regarding use of algorithmic systems?

NO handles complaints of individual citizens when they do not reach a solution with the complaints division of a certain government organisation. NO also has the ability to do their own investigations, they do this when they believe something goes wrong at a significant scale or when they receive a significant number of complaints regarding the same topic. These investigations look into structural causes and what can be done to improve the situation for citizens. A few years ago the report 'De Burger is geen Dataset' (Ombudsvisie) was published, which is NOs most significant publication on the topic of algorithm use by government.

A complicating factor for getting an overview of the problem, is that individual complaints often regard a certain decision that has a negative impact on a citizen, or a certain (absence of an) action by government. Citizens who file these complaints often have no idea what goes on behind the scene, in the 'machine room', and if an algorithm is involved or not. It is likely that an algorithm is involved in a significant number of complaints, but since citizens often don't know this and therefore do not include it in their complaint, it takes a while before NO is able to figure this out.

Is NO still able to discover involvement of algorithms if a citizen doesn't report this?

NO tries to ask more questions and be more mindful in order to get clarity on possible involvement of algorithmic systems, something that was done too little in the past. The approach used to be more pragmatic: if a citizen is helped, the work of NO is finished. There are certain complaints where a pattern can be seen of citizens being subjected to more intense checks, then there are indications that something is happening on the background. However, there are very few complaints in which a citizen is able to explicitly mention the involvement of algorithmic systems.

Do citizens have a better view on the use of algorithmic systems?

The topic of algorithmic systems is more top of mind after the attention for child care benefits scandal, but at the same time many citizens have no idea what data is included in government systems and how often it is shared and copied within the governmental landscape. NO itself is currently working on building up knowledge on the topic more structurally as well.

Asides from processing complaints and doing your own investigations, is NO ever approached by organisations for advice?

Specifically on this topic not insofar as the interviewee is aware. When NO reports on a topic, this creates attention along three lines. First, individual citizens who recognize themselves and want discuss the topic or simply give a positive reaction. Second, and more relevant to NO, intermediaries (e.g. lawyer, debt counsellor) who want to discuss the topic, for instance because their view on issues is slightly different or because they see issues NO didn't report

on. And third, governmental organisations that ask us to think along when they are implementing policy changes.

And NO takes these signals into the view they shape surrounding the topic?

Yes, and NO also occasionally does subsequent research and monitoring, where NO the development of a topic after an initial publication (for example, NO published a report in 2023 about DigID following a publication in 2017 about MIJNoverheid). Complaints are an important indicator, but if it is found the problem is bigger than just these complaints, NO has the ability to open up a contact point (meldpunt) and bring the topic into the spotlight.

What issues regarding algorithm use in execution of law have you seen recently?

We have already seen issues stemming from sharing of data between organisations, where sometimes outdated or no longer applicable data is used. It can impact citizens if something is registered incorrectly somewhere. Citizens are not sufficiently aware of the fact that their data is shared very frequently. In general, people with certain knowledge or contacts will be able to gain some insight into algorithm use, but the typical citizen will probably not be aware of the use of algorithms in processes that impact them.

What would be needed to solve that issue?

You could think of examples such as the Kruispuntbank in Belgium, where there is one central point where citizens can go and get clarity. In the Netherlands, oversight on algorithms is quite fragmented. Another option would be to ensure algorithm registers are linked to significant life events (e.g. reaching adult age, death of a partner), as this would ensure such means are better connected to citizens' world and experiences.

Is the NO also in contact with client councils?

There is definitely contact, also for some investigations. NO also has the 'Raadbaak', a platform where government employees and intermediaries, such as debt counsellors and members of the client councils, can come into contact with each other and discuss current issues. This has thus far not led to signals regarding algorithms.

Is the topic of algorithms still new for NO?

Somewhat. It helps that NO currently has a central research department, where it previously was decentral. That might help with preservation of knowledge. Gathered knowledge, such as knowledge gathered surrounding the Ombudsvisie, doesn't always stick. Researchers focus on new topics or new activities. That's why NO is working to gather knowledge on the topic, and map what their role can be.

What is important for NO to get a better view on issues?

Stories about the experiences of citizens and intermediaries, that is the main way NO approaches it. And, additionally, the ability to interview people in government who have more knowledge about the systems in use. In order to gather experience stories, it is needed that NO translates and traces signals they get. Simply asking an open question about algorithms and opening up a contact point, would lead to loads of non-related signals. It is therefore important to ask a directed question and ask people to contact NO. For instance, questions could related to specific life events that could trigger changes in systems. So, it is important

what questions citizens are asked. And then it is up to NO to trace whether an algorithm could be at play. And in order to be able to trace, it is important we gather knowledge from experts in government.

Is there a role here for executing agencies and supervisory authorities?

In line with the Ombudsvisie, government should communicate clearly what systems they use to reach decisions. This comes back to the principle of motivation of decisions. Supervisory authorities can have a more proactive role, knowing what developments are ongoing, and mapping possible risks of new algorithmic systems that are developed. Sometimes, however, it might be difficult for a government organisation to stop the use or development of a system after it spent time and means developing it.

The Ombudsvisie also mentioned involving citizens in development of systems, does NO get signals as to if this happens and how should this be organised?

Not really, citizens mostly approach NO if they experience negative consequences of not being involved, for instance when it comes to spatial planning (e.g. a new road). When it comes to involving citizens in development of algorithmic systems, the structure of client councils that contact citizens can work. It is necessary to come up with concrete cases for citizens in order for them to make useful suggestions. You could have citizens look at some cases, and ask them what they want to know, what they want to be able to understand.

Is understandability an important factor?

Understandability, or perhaps traceability. Systems can be coupled and decisions can be made based on data gathered from somewhere else. It can be difficult for citizens to then know where to go to fix issues that result from erroneous data, even if they understand what went wrong. There is no central point where data changes are processed for citizens, so citizens have to deal with several organisations. And if no one takes responsibilities for processing changes in all systems, it can take a long time for a citizen to solve an issue.

When NO receives a complaint where it is mentioned or obvious that an algorithm is involved, what is the approach taken?

NO needs to map which organisations are involved and who needs to solve or lead solving the issue a citizen has. If there is discussion regarding a certain system and government organisations do not agree as to who is responsible, we can ask parties such as Logius to give clarity. That regards complaints where it is important to direct solving the issue. If an investigation brings forward something discriminatory or forbidden, then NO will pass judgment in a report, contact a supervisory authority or both, depending on the exact situation. Which authority this is depends on the situation, it can for instance be AP or a ministry.

Is finding the right supervisor in the digital domain difficult for NO and what could help?

Yes, sometimes. What could help for NO is a central contact point where we can go to if an algorithm was used in an wrong or unlawful way. A place where we could signal this issue, and which then takes the issue and ensures it is solved within the government landscape.

D.2 Policy Advisor AI, Big Data and Human Rights at Amnesty Netherlands (I2)

What does your work in this field entail?

The work is done on the intersection of AI, big data and human rights. Research is done into algorithm policy in the Netherlands, especially where this touches upon human rights. Another part of the job entails advocacy and lobbying, in attempts to influence policy of parliament and ministries towards our recommendations, which result from our research. It has to do with human rights, so a part of it is influencing policy, but it also entails overseeing if supervisory authorities fulfil their role and have the right powers, mandate, and knowledge. Amnesty also helps victims of human rights violations attain their rights. The stakeholders in focus are the formal lawmaker, policy makers, supervisory authorities and the "rightholders". It is possible that Amnesty has direct contact with those who are impacted. For instance, in the recent DUO-scandal, we have called for victims to come to us.

Amnesty focussed on profiling and risk models in recent reporting, is that the sole focus or does your team have a broader view?

We have some focus areas that are internationally determined, so for us that would be non-discrimination rights and demonstration rights. Our team looks at both. So for demonstration rights that would be tech risks such as surveillance of protests, for instance using drones or facial recognition, online surveillance of demonstration using OSINT-technology. And then the other focus area entails predictive policing, ethnical profiling and the use of risk profiles for law enforcement, also in the social domain.

One of the focus areas is thus non-discrimination rights, with the childcare benefits scandal and DUO as examples. What is the image you have of this field, and how algorithms are supervised?

Amnesty's view is that the supervision is flawed, on the one hand because supervisory authorities, both internally and externally, do not have the right resources and mandate to execute their functions. And on the other hand it is unclear for organisations how the non-discrimination principle relates to the algorithms they use. So there are issues regarding both supervision and policy, which is what our recommendations focus on. Amnesty recommended that the state conveys the prohibition of discrimination unambiguously, which would have effect in the policing domain but also the social domain. We also called for better mandate, powers, and increased budget for the DCA/AP so supervision can improve. Another recommendation is to make use of random samples instead of profiling as much as possible. What we see in the social domain is that interventions impact those who need help, the use of profiling harms those who already have a weak position in society.

Using random samples wherever possible, you say, but where is the limit? When is it okay to use a risk profile and when should a random sample be used?

The position of amnesty is: no, unless. So, use random samples unless you really have a good reason to use risk profiles. That risk profile should then be tested for compliance with human rights. The use of risk profiles focusses on people without having a concrete and individualised indication that they have done something that is illegal. Individuals are selected based on shared characteristics with a group. Even if they are not directly accused of an illegal fact, these individuals have a higher chance of being subjected to investigations. A disproportionate

amount of investigations within certain groups can be harmful for human rights such as the right to equal treatment, non-discrimination, but also privacy and data security.

In the domain I study, efficiency is used as a positive characteristic of these risk profiles. Furthermore, executing agencies state that algorithms using these profiles can be beneficial for citizens since misuse can be found sooner and thus the amount they have to pay back will be smaller. How then do you see these positive remarks on risk profiles?

The latter is a bit of a spin, there are other things you can do to prevent problems for citizens than using risk profiles. The relation to efficiency is a dominant discourse: that capacity is limited and that you should use it efficiently. But, there is little scientific research that shows that using risk profiles is actually more efficient. So it's not clear if this dominant discourse actually has a solid foundation. When it comes to DUO, for instance, the accuracy when looking at the whole process was quite low. Even though it was believed to be, it was not really that efficient and there was the risk of having to compensate people for lost benefits and scholarships. So there is something to say against this efficiency-discourse.

Relating to that and Amnesty's reports, even though there have been scandals and actions afterwards (e.g. register, DCA), there seems to be some kind of awareness missing within organisations?

The DCA and algorithm register are good steps, but at the same time you see that if its not mandatory to fill in the register or, for instance, perform an IAMA, organisations won't always do it, especially in cases where human rights violations are likely. Our recommendations aim for it to be mandatory wherever decisions made using algorithms have an effect on human rights.

In that case, is this register for the citizen and who should oversee the use of it?

It's not just for the citizen, but also for civil society organisations, journalists, supervisory authorities, and judges to know what happens with algorithms, if there are good safeguards in place, and to what extent use of those algorithms can impact human rights. Amnesty's position is that the DCA should oversee the use of the register. So, the DCA should not just have a coordinating role but also a supervisory one.

Apart from the wider supervisory landscape, there is also the supervision, safeguards, use of algorithms within an organisation. Does Amnesty have a good view on what happens within organisations and does a certain image arise?

The amount of information Amnesty has on organisations internally depends primarily on what the organisations make publicly known. We can of course file open government (WOO) requests if we are doing an investigation. Investigating a specific organisation is always part of domains Amnesty focusses on, namely social security and law enforcement. This is where the organisation sees the biggest risks, also based on prior scandals.

Another recurring theme is the AI Act, and the balance between waiting for this act and taking actions now. How do you see that?

In the case of the police, for instance, Amnesty has already been making recommendations regarding ethnical profiling for the past ten years, and when it comes to risk profiling, the first reports came out several years ago. The AI Act offers some structure when it comes to

fundamental rights, but it also has several exceptions, for instance when it comes to the police. The act was weakened in the final negotiation phase on several areas, so it isn't bulletproof. Our stance is that you don't need to wait for the AI act to ensure human rights are safeguarded, especially because this act won't fully safeguard them.

So when it comes to what needs to happen now, is this about developing new laws and regulations within the Netherlands or about better using what is already there?

The prohibition of discrimination is already there, but the Dutch Government and some supervisory authorities are not that strict in applying it. We hoped that after the KMAR-case the probation would be explained and used more sharply within government, but that hasn't happened. You also see this in the risk profile framework by the Netherlands Institute for Human Rights, which isn't fully in line with jurisprudence: judges stated ethnicity can never be part of a risk profile, the framework states it still can as long as it is one of several factors and it isn't the deciding factor. In this case, the rule is already there, it just needs to be applied according to ECHR jurisprudence.

You mentioned it is sometimes unclear how the non-discrimination principle related to algorithms in use. Is using the judgement in the KMAR-case an example of how a step can be taken to overcome this?

Yes, for instance. Government decided that this judgement only applies to airports, we say it is more of a general principle and that ethnicity should never be used as a criterium. The house of representatives could put down an interpretation of the principle in law, although this wouldn't be needed if cabinet and ministries would interpret it correctly in the first place. Cabinet could take a central place and state the meaning of the principle unambiguously, which might cause underlying governmental organisations to follow. But as long as government doesn't take this strong position, not enough changes, especially because law execution is supported by frameworks and guidelines by government and supervisory authorities.

Gaming the system is used by organisations as a reason to not be as transparent regarding their risk profiles. How do you see this? Is that a valid concern?

From a lawful perspective, it isn't a good reason, citizens have the right to know why they are subjected to checks and investigations. Recent judgements point to this. Gaming the system is a strong discourse in government, but the question is if citizens will actually do it. Transparency should include a detailed description of the system (see Amnesty International (2021), p.31). If this is really not possible, those details should at least be accessible to a supervisory authority.

Does amnesty also focus on more simple, rule-based, systems?

We also look at rule-based systems, as is the case in the DUO-scandal. This is one of the limitations of the AI Act: it specifically focusses on systems that make predictions, whereas there are risks to simple rule-based systems. That is also an example of things you could already do now.

D.3 Leadership role within Enforcement Division at UWV (I3)

What does your work entail?

The department of the respondent is placed within the enforcement division and has final responsibility for algorithm use in the enforcement domain, as well as responsibility for organisation-wide risk policy on abuse and misuse. Models such as CBBS and client application WW are out of scope. There are two types of offences in the SUWI-domain: violating the effort obligation and violating the obligation to inform. The effort obligation regards things citizens need to do to stay eligible for benefits, such as cooperating with reintegration, applying to jobs and not being culpably unemployed. Violation can lead to a measure, which is a temporary decrease in benefits. Violation of this type of obligation is judged within the primary divisions of UWV, the enforcement division oversees how this is done and can support this action with algorithms. Violation of the obligation to inform entails that citizens do not inform UWV of things that can influence the height of benefits, such as income or staying abroad. This type of violation can lead to fines and reclamation of paid benefits. The enforcement division investigates this type themselves.

For which type of violation are risk scans used?

Both until the risk scan stay abroad was discontinued, now solely for effort obligations, so culpable unemployment and the obligation to apply for jobs. The enforcement division does not work with the outcome of these risk scans, that stays within primary divisions. UWV uses model risk management (MRM) policy, which distinguishes different roles and responsibilities over the entire lifecycle of risk scans. The enforcement division has the role of model developer.

The use is within primary divisions, how is communication between them and your group, to see if things don't go wrong?

A lot happens through KPIs, there is a weekly monitoring with a number of standard KPIs. These KPIs are related to several types of bias that can be found in models and the success ratio of an algorithm through a type of confusion matrix. If things threaten to run out of bounds or if there is a degradation in model performance, there is communication with model users. Once every four months the performance of the model is reported on more extensively to the steering groups [coalition risk scan].

Is the 70% signals from risk scans and 30% random signals sent to executing employees (70/30) related to this success ratio?

That has two reasons, one is ensuring that colleagues in the executing roles know that 30% is random and therefore they cannot assume a case is based on a model score and therefore something must be wrong. A second reason is that it allows for continuously creating a control group and an experimental group that can be compared in time, for bias indicators as well as success ratios.

Your group develops risk scans, but how is the logic of a certain risk profile checked initially?

Models can be developed based on inference from data: asking the computer what the most important characteristics are based on the data. They can also be created starting from a hypothesis, that is then supported with data and a model is then built based on this. The

answer to this question relates to the chosen approach: UWV's risk scans are only built based on hypotheses. This starts with development sessions with relatively diverse groups of specialists, where we look at possible characteristics (for instance for staying abroad). The resulting hypotheses are then formulated in data: how can the characteristics be found in data? Here there are several legal checks, then there is desk research, field research, and if everything is validated then it is put into a model. The code is checked by an external validator to see if the resulting code is consistent with the development documentation and DPIAs. The resulting model is then frozen, we don't use models that improve themselves over time. Relatively simple random-forest type models are used, wherein weights of variables and depth of trees are frozen. This makes the logic checkable and explainable. We then check if the performance stays on a certain level and if the output has the groups we also saw in test sets, or if it starts varying.

After the model is frozen, the surroundings of the model can change. Is there any feedback to check for this?

This can be seen in periodical monitoring. Two types of monitoring are used. One is continuous and looks at mentioned performance and bias metrics. Another is a yearly maintenance interval in which we dive deeper and also discuss with executing employees if there are things that can help improve the model to better predict the underlying behaviour we try to predict.

Validation is done by an external party, the MRM-policy mentioned bringing this in-house. Is that still a goal?

That is a matter of taste and thus depends on who you ask. There is something to say for external checks. UWV develops all their models in-house and therefore there is knowledge internally, zo we dare have it checked externally. Perhaps this would be different for parties who do not develop in-house. You then have less control over the risk model itself, and if you then also validate externally, you have very little control.

You mentioned explainability, the MRM-policy mentions client councils, are they involved from the start?

With new models this would be the case, but current models were developed before the MRM-policy came to be. We didn't know then what we know now. The client council wasn't extensively involved from the start, only on main points when the models were taken into production. In a new model for undeclared work we are developing, we involve them sooner. There are many technical and societal developments in the field, which we try to follow as closely as possible. But there are models developed before the societal discussion came about after the childcare benefits scandal came to light. We try to bring measures close to the current reality, but when it relates to the development phase it cannot be corrected anymore.

Algorithms are published in an algorithm register, for whom do you publish in this register?

The ideal thought is that clients of UWV can use it to see how we use their data. The reality is that someone who requests a benefit such as an unemployment benefit can only do this at UWV. Furthermore, they are often more focussed on their personal circumstances (e.g. getting a new job, getting benefits in time) than checking the register frequently. In practice, we mostly get questions based on the register from students and journalists. We often cooperate

with these requests, but perhaps currently the register does not fulfil its intended role for citizens completely.

When it comes to transparency UWV seems to sometimes reason from wanting to prevent strategic actions against enforcement, where does this come from and where is the threshold?

The latter is difficult to define, but the summary is a bit too short. We do not want those who purposely abuse the system to be able to escape enforcement because we give information on our strategy. Research shows that those who violate a rule by accident do not care about this strategy. Those who have the intention to abuse social security schemes will delve into the enforcement system and try to game the system. We try to balance the totality of measures surrounding abuse and misuse. To understand this answer, it is important to realise that when we see fraud risks, an algorithm is the last resort. First, we try to see if its doable for citizens and if the quality of laws is sufficient. Next, we see if our processes are shaped so that the chance of doing things wrong is minimised and the chance of doing things right is maximised. Then, we invest in education and behavioural influencing techniques. The last resort is to use an algorithm to predict abuse or to take repressive measures. The algorithms thus target a small group that purposefully breaks rules and that are not stopped in prior levels, for which it is not needed to publicize an algorithm fully. That differentiation can also be seen in our register: the application scan (sollicitatiescan) is published with all characteristics, and gaming the system is somewhat possible. For the culpable unemployment scan this is trickier, we mention characteristics in categories only. So where the threshold is depends on how we use algorithms, for what domain and for which target group we think it works. We see the public discussion surrounding algorithms, that it can invade someone's privacy, and that every investigation is an investigation. The goal is not to investigate as many people as possible, but rather as few as possible. That is why we take all these measures, with algorithms as a final stop. The leading thought is: what can we do to prevent fraud. But, we shouldn't be naïve when it does still occur.

How does that connect to the idea of UWV to work in a data-driven way? Does it mean an algorithm is not always the outcome?

Yes, and data-driven work can also be used for educating clients, it doesn't always have to lead to investigations. The best example of working in a data-driven way is proactively offering relevant information. Such applications are not threatening, not complicated, and privacy-wise more generic and anonymous. This type of application lies within the directorate client service, which has a similar team as us. When using algorithms in the supervisory domain, we can only supervise individual citizens, not anonymous citizens. It has to be boiled down to a personal level and then we need to check if something is wrong or not. When it comes to client service, we are talking about persona-like applications that are developed in a data-driven way. This is not as one-on-one as an enforcement investigation, so while not without risk, the risk is lower than applications on the supervisory side. It is done by different teams with different policies.

Continuing with the MRM-policy, what does the role of the committee data-ethics look like in practice?

The committee has an expert group, a combination of internal and external employees, and a formal committee chaired by two professors. They are relatively independent from the UWV, and judge if our solutions are in line with our ethical framework, based on the IAMA and the

compass data ethics. They came to be at the same time as the MRM-policy, so it is not as intertwined on paper as in reality. In practice, models are only put into production when the ethical impact assessment is completed.

The committee can give solicited and unsolicited advice. Solicited is thus based on the IAMA, what does unsolicited advice look like?

That is all still new, so it hasn't happened yet for our team.

What is the role of the internal audit division?

The policy has three lines of defence. First is what is organised in the department, such as the four-eyes principle and policy frameworks. Second is our quality department and an external validator. Third is the audit division. The audit division oversees if every player within the system fulfils their role as it was designed. If you assume the policy is good, and everyone does what was written down in policy, then the chance is high that outcomes are also good. The role of the audit division is to prove this. They are at somewhat of a distance and do system supervision.

There are also complaint procedures within UWV. Is there feedback from these procedures?

A lot is linked within UWV, so complaints get handled accordingly, but complaints about an algorithm are few to none. In the four-monthly report, effectivity measures are included, where there is a relation to complaints. So do those who are subjected to a risk scan call more often, can a higher complaint intensity be seen, and is there a higher appeal intensity?

What is still complicated when it comes to organising these processes?

Two things. First and foremost, aiming for moving targets. We try to do things properly and carefully. The MRM-policy was our own initiative. We lean on law and policy, but a lot still has to be filled in. The definition of what good looks like doesn't exist in this domain. Cases where things went wrong have been described, but we are not looking to not do things incorrectly, we want to do things correctly. That is difficult to find. The MRM-policy, for instance, does not guarantee that things will not go wrong. But at the same time it is difficult to say what else needs to happen, because this hasn't been formulated. When it comes to debiasing, for instance, the scientific world does not yet fully agree on it either. And what was good some years ago can be wrong today. And if it is wrong, for instance with the risk scan stay abroad, we are not afraid to admit this, but it does lead to social outrage. It is somewhat being in a split: on one hand there is understanding that algorithms are needed in public services to keep it affordable, but at the same time there is an equally large counterforce that is of the opinion you shouldn't use them at all. So, aiming for moving targets together with a non-exhaustive definition of 'good'. That is very difficult. It is difficult to determine when things are right and wrong, when there is bias or discrimination and when not.

If it is difficult for you to determine this, should it be done elsewhere?

Perhaps it is not possible to determine it properly. In the domain we work in with algorithms we have to prove that something is not there, which is almost impossible. Proving that there is no discrimination in an algorithm, no bias in an algorithm, or in the people that use it. That is complex.

Existing tools such as IAMA, DPIA and different frameworks and guidelines are also not enough to get clarity?

No, because for the discontinued risk scan, we had a DPIA and an IAMA, and still it had to be shut off.

Moving forward, are steps, such as a new framework by BZK and the installation of the DCA, what is needed?

A different nuance is needed publicly when things go wrong, because it makes it hard to work in this domain. I applaud the developments you mention, but there are too many and from different areas. There are many different frameworks, by government actors, privacy groups, commercial groups and so on. They all overlap and give step-by-step plans and checklists. The answer to when it is okay to put an algorithm into production is never given, at least not exhaustively. It would be good, as complicated as it is, to be able to say: if you did these types of tests, it is sufficient. But such a thing does not exist. The insecurity of what is right and wrong makes it difficult to get a grip on what is needed.

In the broader supervisory landscape, is this discussed?

I am mainly in contact with counterparts at other executing agencies, where this is frequently discussed. Policy-wise some things are exchanged. We are confronted with different guidelines from different ministries and agencies, but there is no central control over these different guidelines and frameworks, and on when things are good enough.

To what extent is SZW involved?

Very much so. They gave the orders to develop the algorithms and to put them into production. They were involved in algorithm development and the MRM-policy. Now it is more of an accountability perspective that we involve them. UWV is a ZBO, so there can be some distance between policy maker and executor.

Were they able to help, does SZW have people you can discuss with?

In policy departments the amount of expertise on statistics, algorithms and data is not high. Vice versa, the expertise on policy making in my team is also low, so we work with legal advisors and such. In general, there is work to be done on the knowledge on AI, data, and so on, also in politics.

The lack of knowledge can also be seen in the reaction to the DUO-case. The focus was on the algorithm, but a big problem was how it was used by employees. This is difficult in the public debate, that it is easy to talk down on algorithms. The alternative is manual checks by employees, but bias does not disappear: humans are also biased.

D.4 Researcher at Court of Audit (I4)

What does the work of the Court of Audit regarding algorithms entail?

The Court of Audit [ARK] is an external and independent auditor of the Dutch government, which does risk-based investigations. Algorithms is one of the subjects that is researched since a few years because the risks are high but supervision wasn't properly arranged. ARK started out with the algorithm assessment framework (toetsingskader) to see how algorithms should be investigated. This framework includes relevant law and policy and formulated risks of algorithm use. Since it was introduced, ARK has used it to test algorithms. This is done because ARK sees there is work to be done to control risks, and systems are currently not investigated enough.

How are these investigations using the framework done? How are algorithms selected for instance and what is the relation to other investigations?

Algoritmes getoetst was the first investigation where nine algorithms were looked at. There, ARK selected the algorithms based on risk of use. Next, the framework was used and together with different disciplines the control over risks was looked at. Then, internally the measures made to control risks are discussed for adequacy. The first steps are done together with the organisation we are investigating, then ARK makes a final judgement. Since last year investigations into algorithms are done as a part of the yearly accountability report (verantwoordingsonderzoek), which means its researched more as part of processes. Together with ARK's departmental research team, the process that an algorithm is part of is looked at, in order to see how the algorithm behaves in its context. This enables the team to see the consequences of not adhering to the assessment framework. In doing this, ARK is able to judge whether there are areas of attention or even deficiencies.

Some years ago, ARK published an overview of algorithms in use within government. Is this overview then used to select algorithms to research and how does ARK get an image of where algorithms are used in the first place?

That remains difficult. That first research was not a comprehensive overview, but rather a research into whether the departments have insight into algorithm usage themselves. The conclusion was that this insight was insufficient. The research can be a source to select algorithms, but ARK also keeps its ears and eyes open and looks at supervisory authorities, governmental organisations, researchers within ARK, and the news, to see where else algorithms are used. At the same time, insight into where algorithms are used is still insufficient and remains a significant challenge for supervisory activities.

The assessment framework mentions the possibility to contact ARK for suggestions. Was the framework built in cooperation with other parties and have changes been made since?

It was developed in cooperation with others. Several expert groups were organised, for instance with ADR and some accountancy organisations. Discussions were had with some ministries as to what the most important risks of algorithm use are. The main foundation was existing law and policy that was bundled in order to be able to make judgements on what is important. The framework has recently been renewed, it remains a constantly developing artefact. At the same time, ARK is not the party that sets norms or makes policy, that is BZK in this case. So when BZK publishes their new framework, those are the norms that ARK will test

for. That is not to say ARK's current framework will not be valid anymore. When it comes to BZK's framework, ARK tries to advise based on insights that have been gathered over recent years, but is not a co-developer. That is not ARK's role.

In the report Algoritmes getoetst, six out of nine algorithms that ARK looked at were flagged. How were the responses on this report?

The reactions vary, as can also be seen by the response by government. Some organisations are happy that the test has been done and gathered points of improvement to work on. Other organisations were less happy with the report. However, several organisations went to work with the recommendations that were made to change internal processes. And that is why the test was done in the first place.

After stating an algorithm does not adhere to the framework, is there successive action later on or is it to be dealt with internally in the concerning organisation?

The organisation is meant to work on improvements by themselves, but of course this is something that is monitored on. Such things are not solved quickly, so the idea is to periodically check up on the status of improvements.

The assessment framework encompasses several dimensions, some relate more to (impact on) citizens, others more on the organisation and effectiveness. Is there a central goal to the assessment framework?

For ARK the largest goal is that algorithms are used responsibly, which related to several different aspects. So the framework is made to judge whether algorithms are used responsibly.

What is responsible?

Safe, trustworthy, honest, and so on.

That seems to focus not just on 'classic' court of audit themes such as effectiveness and efficiency, but also have more attention for citizens specifically.

That is something ARK has had more attention for recently, and it is part of ARKs strategy to look at implications for citizens and companies and focus our investigations on that. Responsible use has to do with legitimacy, but also if it is honest, responsible and safe for citizens and companies.

The framework of ARK is used by ARK to investigate algorithms, but is also meant to be used by organisations themselves and by other supervisory bodies. Does ARK help in this process? What is the image that arises from such contact?

ARK has plenty of contact with executing bodies themselves, but also supervisory bodies or municipal courts of audit that used the framework and want to exchange views. Executing agencies are still searching for a normative framework that gives guidance as to when they are doing things right. The normative framework gives some insight into the risks. But if all these risks are controlled, the framework doesn't state if an organisation is doing things 'right', so they are still looking for that. There has been contact with municipalities and courts of audit when they were actually investigating an algorithm, mostly about the interpretation of the framework and problems they ran into. It is good to exchange experiences together.

Is the question of whether they do things in the right way focussed on proving this, or actually on controlling risks?

The latter. Organisations have the will to responsibly use algorithms. They see algorithms are necessary and they want to use them in a good way. Currently, however, they are still searching for: what is good enough?

Coming back to the view organisations have on their algorithms, do you think that has improved since the initial overview ARK made?

That is difficult to say, because we see very little information on this openly. When looking at the algorithm register or the information given to the chamber of representatives, it is not necessarily better. But it is improving. The algorithm register is there and it's slowly being filled. In monitoring conversations ARK has we see that organisations are busy trying to gather further insight. So it is improving, but we're not there yet.

Regarding the register, have you compared the register to the image you have, for instance based on the overview? Would ARK like to use the register, or is this perhaps more for citizens or other organisations?

Yes, the comparison is made of course. But if you see that big organisations only register one algorithm, then you already know there is more going on behind the scenes. The difficult thing about the register is that it has various target audiences. It is good if, in the end, the chamber of representatives, but also citizens are informed about where algorithms are used and what information is processed in the process. Of course, ARK will also use this for investigations.

One of the things ARK looks at is transparency. Some organisations are somewhat reserved in for instance using the algorithm register, for reasons such as gaming the system. How can this take shape?

The algorithm register is a tool and not a goal in and of itself. Transparency can be achieved in several ways, one of those is that citizens have a right to know how decisions were made. If they don't get a subsidy, they should be able to determine what information was used to come to that decision. Citizens have a right to this, so as an organisation you should be clear when it comes to that: an algorithm was used and this information was used. How you give this clarity? This can be using the register, but also other ways. It seems logical that organisations don't publish all models from a to z, but they should find a fitting way to be transparent.

In the broader supervisory landscape, things are changing. There will be the DCA on top of existing supervisory authorities. How do you see the changes and the role of ARK in this movement?

As independent auditor, ARK isn't directly part of the supervisory landscape. We hope that this landscape is sufficient without us, and we do risk-based investigations in addition to it. In the landscape, it is good that there will be an extra supervisory authority, who can have an integral vision. It is a challenge when it comes to algorithms that several disciplines come together: you need the privacy expert, but also the data specialists, IT-controllers and ethical and legal specialists. That is a big challenge for supervision on algorithms and AI for the coming time.

If ARK is not directly part of the supervisory landscape, then what is your role in cooperations in this landscape?

It is important to be clear about the role of ARK. It is an external auditor, High Institute of the State, with a legal task to audit the government. In this case also specifically when it comes to algorithms. That means we keep an eye on the workings of the supervisory landscape and try to strengthen it as much as possible with knowledge and expertise we built on the topic recently: inspire and help supervisory bodies and inspectorates, and exchange thoughts, also in order to see where challenges lay in coming years.

In bringing together laws and regulation in the framework, did you see blind spots where something needs to happen, for instance additional law or changes to the supervisory landscape?

A lot is already put into law and policy, for instance when it comes to privacy (GDPR) and IT (e.g. BIO, Baselines Informatiebeveiliging Overheid), things that everyone should adhere to. Now it is important that this happens integrally. That is an important step: looking at problems integrally and not fixing one issue in a step-by-step way with a singular focus.

Apart from the assessment framework there are several other frameworks. Is it possible to bring all things together in order to do things integrally?

When looking at the several frameworks and other instrument, you can see they all have the same purpose. Sometimes there are duplicate parts of these different tools. The IAMA has components such as bias, human rights, but also GDPR-like parts and the BIO. If you fill in an IAMA, perhaps part of the assessment framework is also covered, as is the framework of the ADR, which is quite similar to that of ARK. The contents of the AI Act, at a higher level, do not differ all that much from it either. So it can be brought together.

It is important as an organisation to make an integral trade-off. Look at the opportunities, but also at the risks. See what you need to need to consider and if, in the end, it is responsible to use a system.

D.5 Consultant at National IT Guild (RIG) (I5)

What does your work entail?

The RIG is part of BZK and has a group of engineers that are deployed throughout the national government on temporary or project basis. There is a group of about eighty people, thirty of which focus on data and AI. Respondent has done several projects for different governmental organisations: when I started around six years ago as a data scientist, I thought I would train models and such. But that is not what was needed, organisations had good data scientists. The trouble was getting these models from the experimental phase, in data and AI labs, into production. This was due to legal, ethical, organisational and technical factors, that made this step difficult. That is what sprung the start of the RIG.

How has that developed, is that step still difficult?

It can still be difficult. Technically, a lot has improved. There used to be no real technical infrastructure. You had to explain to the IT department what AI is and how it is (dis)similar to regular software development, and why certain provisions are needed. Data scientists had to be told that at the end of the day, it is just code, so you have to do version control and such practices, already common in IT. Now, the focus is more on the organisational, legal, and ethical components: how do you organise these things properly? And, where we used to deal with the project team and the things they had to do, the realisation is starting to set in that it requires broader organisational changes. It is in the core of the organisation. It is not just about the project team or the IT department, but also concerns the layers above them. That is a big change.

How do you gain insight into how this should be organised organisation-wide for the different organisations you work for?

It matters how you get involved. This used to be bottom-up, via a project team or lab, and then it was a lot of work to get the point across that it is not just a project. You need to organise governance, make policy, and such, at the highest level. Because of two things, we see that we currently get involved more through the higher levels, more top down. The first reason is the accidents that have happened. Unfortunately, we often work for organisations that have been confronted negatively with risks of AI. That shock made them aware that it is not something you can simply purchase, or make in a lab, and then implement. It is something you need to think about carefully as an organisation, you need to prepare for it, organise governance, risk management, educate people, and so on. That realisation comes with a shock sometimes. The second reason is upcoming actions and regulation such as the DCA or the AI Act. That is another big change. People realise a lot is happening and action needs to be taken to prepare for it. In the end, you need a sponsor at the highest level. Involvement from top levels is needed, and it prevents needing to work your way up. What is needed within government is leadership. There is work to do at the executing level, on the work floor, but the realisation that things need to be done carefully is present within data science teams. It is important that this is felt organisation-wide, and for that it is important higher levels and leadership are involved.

And then, how do you decide what to focus on among all the current laws, actions, frameworks, guidelines and so on?

It is important to offer perspective. A lot of organisations have a lot coming at them and are overwhelmed. It is important to find out: where are you now, what is your level of maturity? What is your ambition, how fast do you want to get there? Then we can lay out a path and see what the first steps are. An interesting example is ethical committees, which are seen as an important thing at the moment. That can be good, but for organisations that have never worked with data ethics and responsible use of AI, it is not the logical place to start. It might be a good sign towards the outside world or higher management, but it doesn't fit the maturity level of the organisation. When you look at the AI Act, a lot of talk has been about the definition of AI, and in extreme cases it stalls progress. This is unfortunate. The AI Act is about a risk-based approach. You can use a risk-based approach for your AI governance and risk management, even if you do not know the exact definitions and applicability of the AI Act. So I try to decouple these things: as an organisation, you want to responsibly use data and AI, and you need to take certain actions in order to do this. You need to make an inventory of where you use algorithms and AI. You then need some policy on how to deal with AI, you need certain roles in place. You need to do risk management, so you need to do impact assessments. And not in the traditional sense of: what risks are there for the organisation, but: what does it mean for the outside world, for the citizen? And then you need to couple actions to the outcomes of these assessments and monitor for it, the latter is important but there is not a lot of attention for it. If you do these things, no matter what exactly is in the AI Act, you are already far along. It is important to make it tangible. The top level, the directors, are aware that AI offers opportunities but also risks. And at the bottom levels, those project teams, this awareness is also present. But there is a gap in between them. And that should be filled in with governance, risk management, and so on.

Is that difficult, the translation of certain risks, rights issues, to what is happening?

The point is, when speaking about things in an abstract way, you end up with statements that everybody agrees with. So for instance in the AI Act or policy of the organisation it will state that there should be a human in the loop, you should not discriminate, you need to respect privacy. Those are high level statements everyone agrees with. Once you look at a specific use case, things become difficult, as you need to choose between certain values. Organisations are not organised so that they can do this. Technical teams often cannot recognize these ethical and political questions. Frequently things are political, and perhaps choices you shouldn't make as a project team, but higher up. And another thing that is difficult at an abstract level is quantifying things in order to monitor and optimise for them. It can be difficult, also for managers, to state how you quantify a certain value, such as fairness, for a certain context, and what acceptable margins are. This is also difficult in the National Standards Body, that needs to translate certain abstract demand of the AI Act into technical and process specifications. This way of working was used in product safety previously, where it is perhaps easier to agree based on a shared idea of safety. Now, we are not just talking about physical safety but also a broad array of fundamental rights, for which we have to find technical specifications. That is difficult.

Some things are perhaps easier to quantify than others. How do you deal with the remaining things that cannot be quantified easily?

That is difficult in my work right now. It is very good that we try to quantify and specify things such as transparency, explainability, non-discrimination. But we should also be aware that some things remain. And for those things it is important to have the right culture, have the right conversations, educate people, peer-review, coach people, and bring different organisations together to discuss. That is very important, but perhaps does not get enough attention. We also need new roles in organisations to facilitate these discussions ánd to document them. That is also very important. We have a lot of IAMA-sessions and so on, that is good, but they should be documented, also for the follow-up. That is the assurance part, which is very important. That can very well be done in standards, there will be rules for it and a supervisory authority, as you can say things are in order, but it is important to be able to show this. That can all be fixed with laws, regulations, standards. The part about operationalising fundamental values and monitoring them, that is very difficult. My mental model is: we want to reap the benefits of AI, so innovation is important but in a responsible, good way, so we can trust as citizens that it is safe. That is what we want, and on one hand we centralised this in a top-down manner with law and policy. Then we have institutes making standards, stakeholders coming together and trying to find consensus on what the characteristics of a system need to be in order to safeguard safety and quality. That is more inside out. And then we also have projects such as algoprudence, more decentral and bottom up. Those first two are necessarily somewhat abstract and not very context specific. And for the remaining things we talked about, it is important to look at the context. Different contexts warrant different trade-offs. Perhaps ideas such as algoprudence can help here, having conversations about certain cases and deciding bottom-up how it should be dealt with. And this bottom up action can also inform policy and standards that are created higher up. So top-down, inside-out, bottom-up.

Perhaps somewhat overlapping with algoprudence, you mentioned you make use of use cases?

Yes, because these trade-offs and ethical dilemmas can be very abstract. So using a case can make it concrete. We use both cases where something went wrong and new ones. The latter are the best, as they make people enthusiastic and spark energy. A case where something went wrong and that caused trouble is never easy, but can also be used to learn from, for example DUO and Rotterdam cases. It is disappointing to see these organisations get a lot of backlash, even though they dealt with mistakes in a good and open way. We can learn a lot from these cases in this field of work, because it makes things concrete. The backlash can prevent other organisations to be as open about their struggles and mistakes, whereas these stories are very important. In the DUO-case, for instance, different reports had different views on the certainty with which bias could be identified, and whether bias was present in the algorithm itself or in the decisions made using the algorithm. However, there is no room for this, things have to be very binary: right or wrong.

Can agreement be reached as to what organisations have to adhere to in that case?

You need to accept that there will not be a full answer to this. So it is about coming together and discussing what is acceptable for society. It touches on democracy here. You need to solve conflicts, and come to a certain agreement. That can change over the years as different things are found important or if we learned new things or so on. But dealing with uncertainty is very difficult for organisations.

So uncertainty in what it means and what they need to adhere to?

Yes and the legal, ethical, political values that you need and that you translate to technical specifications. That is a struggle. You can have a process surrounding algorithms, their development and use, that has monitoring, is careful, and has checks and balances, and still make an algorithm that is wrong (e.g. discriminates). It can still interpret certain democratic values in a way that we do not deem acceptable. I am curious to see who will assess and audit that and how, but also worry if there are enough people and funds for it.

That would be an outside view. For organisations you are an outside view right?

Yes, but in an organisations, it is difficult to answer questions such as: does this algorithm discriminate. It is also not a question that can be answered by a technical person.

Who should answer it?

I would like to have a multidisciplinary team with a statistician, a data scientist, but also a behavioural expert, a communication expert, an ethicist, a legal advisor. And then also people with knowledge of different domains and representatives of those who are affected by an algorithm in a specific context. And then you can properly answer such a question. We as technicians think about certain things, such as equality or non-discriminations, and try to understand what it means in this context in means of outcomes. The different metrics you can think of and what is deemed acceptable and what not. The metrics and the numbers are very important for the discussion, but it can also hide things, for which qualitative research is very important. That is sometimes missing. At some point you want to do a root cause analysis to understand certain outcomes of a model. Perhaps there is a plausible explanation, and then you can think of interventions. But you need different perspectives to find these explanations and think of interventions. That is a next step, to work in those teams. Working in such teams is very interesting and important, and it needs to be scalable, for which a risk-based approach can work. But at the same time, it is expensive, takes time, and the people are not readily available.

Is the basic question of: should we want to use this, why do we use this, ever discussed?

It happens, more often, so that is good. But you also see sunk cost bias: it is difficult for organisations to stop projects after having invested a lot of time, money, and effort into it. That is another interesting form of bias that is difficult to quantify. Sometimes you are asked to look at bias in an algorithm, but you can never decouple this from effectiveness of an algorithm. If an algorithm is not effective, then the answer to proportionality is no anyways, bias or no bias. Sometimes a baseline is not measured, and the idea of subsidiarity: can we reach the same effects in a less invasive way, is overlooked.

If an algorithm has been used for a while and becomes part of the system, is it even possible to have a good baseline?

That is why it is important to think about this in due time. And, this touches on another point: there are reasons why using machine learning models in government is more difficult than in other sectors. You directly influence citizens, and can change things you cannot change back. It is difficult to see how things would have been if we would have done things differently. That can also make it difficult to calculate certain metrics, the confusion matrix is not always complete for instance. And then it is difficult to say something about effectiveness.

D.6 Policy Advisor at Netherlands Institute for Human Rights (CRM) (I6)

What does your work entail?

Respondent is a policy advisor, specialised in discrimination law. CRM is a supervising institute focused on human rights, with a mandate that consists broadly of two parts. It holds general oversight over human rights in accordance with the UN, with the goal of protecting human rights nationally. CRM was therefore appointed by law as supervisor, and thus has a legal basis. The institute is also an equality body in accordance with EU law, overseeing equal treatment directives, and therefore it has the ability to publish judgments. Complaints can be filed to CRM if one thinks they are discriminated against. These complaints are then judged in a court-like setting, wherein parties are obliged to appear. The judgement is not binding but often adhered to. Respondent works in the investigative branch of the institute, although there are connections and knowledge exchange between the branches. This branch, however, is not involved in judgements in cases. The investigative branch has a programme focused on digitalisation and human rights, and respondent is especially focused on discriminatory aspects of digital development. CRM is an overall supervisor for human rights, but these programmes offer a special focus, within which there are several subthemes, such as digital accessibility and digital literacy.

In these themes, how do you gather necessary information? Some supervisors can request information, how does this work for the Institute?

We are authorised to do investigations at the scene. When it came to the childcare benefits scandal this was used for instance, because there were many requests for judgements, so there needed to be an overarching investigation to see what happened at the Tax Office. This investigative capacity hasn't been used for the digitalisation programme. We worked with contacts within the field and within the scientific field to gather information on a voluntary basis. And we have our own employees that are experts on the matter. There are general scientific technical insights and we have our own expertise on law. In the end it is about bringing it together and seeing what a certain technology means for law: does it mean something new and is it a threat? That is the expertise of the institute, with a broad and deep knowledge on human rights, which we can apply to new technologies. It is necessary to gain some technical knowledge first in order to do this, for which a specific programme is a good way of working, as it bundles it and brings about a focus.

Is the crossover between law and human rights obvious in many cases, or a difficult search?

For the institute it can be obvious, due to the knowledge position we are currently in. In the field we see that there is a gap between people that have knowledge of the law and organisations that have technical expertise, but are not always aware of the risks for human rights. In organisations, we see that those who work with data and data applications, the data scientists, find it difficult to assess this. They have knowledge of data, statistics and bias, but that is different from prejudice in a legal and societal sense, which is not statistically measurable. Discrimination, in the end, is a judgement for which you need societal sensitivity, moral judgement, legal knowledge and legal judgement. This can be difficult for the more technically trained specialists. It is our job to try to bring it closer together.

The technical experts have trouble with making these judgements, are there other places in organisations where this is better?

Things are developing quickly, partly due to publications and the childcare benefits scandal, wherein the role of algorithms was perhaps somewhat limited. The attention for the topic has increased massively, which is also visible in politics and investigative journalism. The focus is quickly on the algorithm an organisation used. If you look at cases, DUO for instance, the problem is often broader than just the algorithm. The public probably thinks of a complicated, self-learning AI system when hearing 'algorithm', but in the DUO case it boils down to: what criteria are you allowed to use? This could have been a discussion before the information age. The attention makes it easy to problematise algorithms, but if that gets to the core of the issue is questionable. Within organisations, knowledge on discrimination law is felt to be insufficient. Making judgements on discrimination of digital systems is difficult within these organisations, and they perhaps depend on others. Even if technical knowledge is present, how and where you can assess these issues is difficult. For some it can be an eye opener that even if certain systems function, for instance to find fraud, they are not automatically allowed. There are more possibilities nowadays to make data applications, so they are also used more often. There used to be risk predictions and risk taxations, based for instance on a team's experience. Now there are data applications, new ways to make connections, but the question is no different than it used to be: are you allowed to investigate certain people more frequently based on an educated estimation, and what consequences can this have for specific groups?

This ties in with the paper CRM published, that spoke about transparency when algorithms are involved. Are existing laws sufficient?

We do not see transparency as a standalone human right, but it is needed with the increase in algorithm use. Discrimination is always hard to prove, you get denied for something or you get investigated, but you do not know why or how they got to you. With the increased use of algorithms people still don't know this. You often only receive a letter with a decision, and only know if an organisation selectively searched if you know everyone that also got a letter, and everyone that didn't. It is new that this happens more often. If algorithms form a risk, it is important to know an algorithm was used in the first place, for which a transparency obligation is needed. This is needed for two reasons: one, people need to be able to defend themselves. This ís a human right: right to a fair trial. Especially when algorithms are used, there is a risk of discrimination. And two, for science and investigative journalism it is important that there is transparency, so there are extra eyes on these algorithms. The systems are complex and not every citizen has the knowledge to understand them and their outcomes, transparency can enable a kind of counterpower. It isn't the ultimate solution, but a necessary condition for people to defend themselves against possible discrimination.

The paper also mentions the algorithm register. Is that the most important way to be transparent, or an example?

It is an example of how you can do it. The point made in the paper is: you need to be transparent, whichever way you do it. Given that we already have a register, this can be something to use. Citizens are not waiting for complete technical transparency, so this can be counterproductive. The register can play a part: if this is in order and obligated, you or your lawyer can for instance use a characteristic of an algorithm that was used to find what has happened. It is the choice of managers, and not us, how transparency takes place, as long as it takes place. The register can be a logical instrument. We also share scepsis about it, however: that it is not obligated and that it is incomplete, especially where there is tension. We also

know it is not possible to have a working and complete register from one day to another, so the current register is not completely suitable for this task, but it can be a direction to go in.

The paper also mentions fraud detecting algorithms and different transparency requirements, correct?

There can be special reasons not to release certain information. This exception can also be found in the AI Act. Some things can be investigative information, and releasing it can give perpetrators more knowledge than is wanted. Not all citizens, but those who want to commit fraud could try to game the system. Those are realistic considerations. We don't say, don't release anything, but rather that if certain parameters are sensitive, then there should be a way to not disclose those. However, in that case, an independent organisation should be able to assess whether this sensitive information can cause discrimination. If this isn't established, it stays hidden and it becomes impossible for an individual to prove discrimination, which means there is no legal protection.

You don't mention what kind of organisation this should be, do you have thoughts on this?

That is an implementation issue that we haven't answered, it is about the principle in the first place. Who should do this can also depend on the sector, but it should happen and be fitting, executable and independent.

Coming back to trouble organisations can have with discrimination laws, and the dependency on another party. Who can better explain how they should understand these rights?

We try to play a role in this. However, executing agencies mention they simply do not have enough expertise on this topic. You cannot outsource everything, but should try to gain knowledge within your own organisations first. You're responsible for the instruments you use, and if you are not able to assess them on human rights, that is an issue in itself. It means you're not able to take responsibility for your own actions. So that is the base, and then it is important knowledge is spread. We already try to play a role in our educational role on human rights. But, it would also be good if there are good methods to assess algorithms throughout their lifecycle. And, after the fact, if there are more judgements on what is fundamentally allowed and what not. That could speed up knowledge development. There haven't been many cases concerning algorithms at the courts or at CvRM, although it is starting to show up.

The Institute itself has published an assessment framework on risk profiles, how was that received?

Reactions varied, some are happy and others still need more explanation. The framework holds norms that you need to assess for, such as legitimacy, proportionality, subsidiarity. Some government employees still find them hard to apply. It is a good base, but it is imaginable that us or others will develop specific instruments to explain how to actually assess. There are more general instruments, but specific ones are needed for the entire lifecycle. CRM could play a role here and explain in smaller steps what organisations need to look out for. The current assessment framework is a good step and there will be an update, and perhaps it'll become a living document that is expanded with new jurisprudence.

Complaints are starting to arrive, but are still few. How does the Institute know an algorithm is involved when a complaint is filed?

There haven't been many complaints, so it is difficult to discern a recurring trend. There is a front office that looks whether a complaint can be taken on, based on if it falls under laws we can judge on. On first instance, suspected involvement of an algorithm comes from signals of those filing a complaint or their legal aid, for instance when they got this impression from documentation. Sometimes the suspicion comes later on during the procedure, for instance from reports. It varies.

Does CRM ask clarifying questions in order to find out or does it completely come from the citizen?

The initiative lies with those filing the complaint. There are also anti-discrimination points, each municipality needs to have one, that can help file a complaint at the Institute. We try to help educate employees of these points. But it is still a point of concern, which is exactly why we argue for more transparency. When they are not told, there are few citizens that have enough knowledge of a municipal organisation or executing agency to reason what instruments have been used. Transparency is a first step, for which an algorithm register can help, but it has to be mentioned and be complete.

We spoke about what needs to happen within organisations: translating laws to actions, transparency, knowledge development. When looking at the supervisory landscape, what are necessary changes that need to happen?

In the first place that supervision needs to be organised. This sounds simple, but currently there is no functioning supervision over algorithms. There is a role for CRM here, because of the AI Act but also because general supervision needs to be organised. It is necessary that supervision is organised, as we see the citizen cannot do it own their own, because in principle a citizen only knows their own case. Legal protection, also at CRM is currently bottom-up: a citizen has to file a complaint and prove they have been discriminated against. In a purely bottom-up fashion, you won't fix the issue. Supervision will help prevent discrimination in a top-down way, which is key. The AI Act will help, under which CRM plays a role alongside market authorities. The AI Act works on a product level: which products can be offered and which cannot? That is something else than supervising whether products are altered after they are in use, or whether the context changes and therefore the functioning of the product changes and becomes discriminatory. There needs to be something for those points in order to make the supervisory landscape sufficient. The first steps, however, are already a large amount of work. The supervision over risk profiles that are deemed sensitive and therefore cannot be made public also needs to be organised, which is important. This applies to algorithms and AI, but also to non-digital risk profiles. There is internal supervision to some extent, but unfortunately we see that internal signals are not always picked up. External supervision can help here.

D.7 Employee of National Client Council (LCR) (I7)

What is your involvement in this topic?

Algorithms are not easily accessible when you're on the outside of an organisation. Respondent gained experience with algorithms back when they were used for research and analyses. Data could be used to research effectiveness of policy without having to approach citizens, and coupling of data from different sources made these types of analyses easier. This way of working was built together with public organisations and research institutes, where it was seen that data and algorithms could be used to do explanatory research. At the same time, this gave insight into the mistakes that could be made when using algorithms. Later, algorithms became more widely applied. Where it was first used to see where policy wasn't effective and thus where people could use help, now these systems were also applied to sanctioning. The Bulgarian fraud scandal led to a larger focus on fraud, after which several governmental organisations started developing algorithms on their own. That led to systems such as SyRI, a system that the LCR was involved in investigating. A lot of multivariate analyses can be done, but rules surrounding it should be in place to see: is it fair that someone is suspected of fraud, based on what is this suspicion raised, is the sanction proportionate, is it motivated correctly, and how was information gathered? With SyRI, the LCR had her doubts, because there was somewhat of a drag net. Data was used in ways that made people suspects, but the relation between findings and suspicion was unclear.

Another interesting case was the Wijdemeren-case, a municipality that used a private company for their fraud enforcement. This company coupled data to distil a profile that could be used, and sent signals to the enforcement division of the municipality. The municipality thought they were transparent and didn't know what they were doing wrong. LCR was part of this discussion, and there was the wish to be able to assess algorithms, because there is a certain selection made in building the algorithm, in the variables taken into account and how they are structured. If you look into this, you can assess whether valid assumptions were made about what the model could predict. Intellectual property was brought up and prevented directly assessing the algorithm, after which a trusted third party was taken on. The data turned out to be unstable, after which this method for looking for signals was stopped. Even though the municipality wanted to do it right, it still ended up being problematic. Working with a trusted third party was insightful. This is an example where an algorithm is used and supervision is difficult. Fraud detection happens, but you should do it in a proper way and models used on citizens should be tested. Expertise on this should be organised around the systems somehow. The supervision should be somewhere. AP will do something, which can be seen as a trusted third party. However, they are also seen as enforcers, so that can be difficult.

I also saw a research project where algorithms were to be used to figure out the relation between medical issues and sick leave frequency. From a research perspective this can be a valid question, but the impact on citizens can be immense, for instance when it comes to selection effects with insurers and employers. It was approved by an ethical committee, but had not been tested for possible effects of outcomes on society. In the end there was a collection of interest groups that protested against it and the guiding committee of the research put a stop to it. Even though an ethical committee assessed the research, it was not assessed by societal groups such as patient organisations, employees, employers and so on.

Such research questions and algorithms should be assessed by groups grounded in society. Different groups should be involved, because impacts of certain data are not always foreseen.

Does LCR currently get involved when SVB or UWV develops something?

It was proposed, but it does not happen enough. They have to know certain expertise is present and probably do not expect LCR has this, but we happen to have it. Influence is still applied through guiding committees LCR takes part in. If LCR had to assess all the algorithms of an executing agency, we would not have enough manpower. But for certain algorithms it would be good to have an assessing committee where this is taken into account. LCR is involved when incidents happen. The incident surrounding the risk scan stay abroad is another example of this: cookies were used to see if someone logged in abroad, which was connected to possible stay abroad, which meant they could not apply for jobs as required. Another example of an algorithm which involved a risk profile that was used for enforcement. For the average citizen or for LCR this is not traceable, unless citizens get a signal that they are being looked into somehow. As an individual that is hard to find out. Sometimes we see that decisions from executing agencies are delayed even though they should be made, which leads us to think that an individual is under the suspicion of fraud. Executing agencies are not transparent about this, and citizens do not see it. They put in effort to ensure a certain decision gets made, make legal costs and are under stress, but do not see that it is all because they are suspected of fraud. The lack of transparency leads to frustrations, and we can only retrace this indirectly. We talk to these people and they share their story. We only started seeing this as a missing link in cases recently.

Is it thus difficult for LCR to gain insight into how algorithms are used within executing agencies?

They are not transparent about it. They state it is strategic information and sharing it can lead to an information advantage and possibly gaming the system. So it is kept a secret. And if things are kept a secret, because of intellectual property or because of strategic information, it makes things untransparent and thus vulnerable for errors. You could talk about the systems that are used in order to prevent errors, but that is not done. Sometimes certain errors are made, even though they could be easily spotted by someone with the right knowledge. The impact on citizens is potentially very large. The cumulative knowledge about what is being undertaken and how this should be organised is missing. Fraud exists, even if it is just a few percent of all citizens. You still have to do something about it, even LCR stands behind it. But because this happens without proper supervision, and because checks and balances are still not there, who is to say mistakes made in the past will not be made again? A dialogue should be had on this, with the right knowledge present, including those in an executing role. This should not be done with certain interests in mind, but rather as a way to look at things together and see what is happening and what it should look like. In a past role, I was a person that was called in to see what was underneath the data. Statisticians would make algorithms and have certain findings, then this expertise was called in to see what underlying explanations were, what plausible assumptions were, and if they could be cross-validated with other findings. I am not involved in these things now. The only thing I see is the incidents, such as the benefits fraud, and you see things got out of hand and the process was not organised in the right way, with the right people. The focus on fraud and enforcing rules led to tunnel vision in order to get results. By being more careful in the processes you use, you hope that tunnel vision is lost. But I do not have the impression this is happening at scale, the trust is still not

there. Organisations are responsible for it, and they might have good technical monitoring, but they do not see the societal implications fully, nor what is beneath the variables and if what they assume is plausible. That extra test is currently missing, and we can improve by looking more broadly than just the technical checks and not just involving statisticians but also interest groups and such. The examples I gave show when the wrong judgements were made. Even if you can research into depth using data, the question is if it is societally desirable. There are boundaries to what we should want to know in order to not bring individuals at risk that are in a collective arrangement (e.g. insurance). That is a new insight that we don't see enough in models that are made. The discussion is often about the technical side, and if it is allowed from an ethical stance, but even if from that perspective it is allowed and it is technically possible, should we want it? That question is not asked.

And even in ethical committees and such, this question is not taken into account?

I gave the example of research that was passed by the ethical committee, even though al societal groups were against it. That is because it is judged in a different manner. Clients and employees, looking at it from their perspective, use different criteria than an ethical committee. Then there is also the risk of the data and conclusions of approved research being used to build algorithms that are then used in selection processes, over which there is little control. So simply doing a research can lead to the insights being used for other purposes. An ethical committee does not control this: they look at their own organisation, not what others might do with certain information. If ethical committees have an internal focus and are not focused on effects on society and consequences for others, there is the potential for damage. That is the biggest objection when looking at ethical committees, they do not take those other factors into account sufficiently.

Do you see differences between organisations, for example SVB and UWV?

They try to do things properly, I am not sure if they do. A new piece of legislation was just sent to the chamber of representatives about proactive service. It is about proactively coupling data within the SUWI-domain, in order to find out who does not use certain benefits schemes, so these citizens can proactively be approached. In doing this, poverty due to non-use is prevented. But, this is the same thing as sanctioning, but reversed: positive. The same type of data is used. Proponents state that we all find it important to proactively help citizens. However, when this is done somewhere in a department of an organisation, and algorithms are put into action there without supervision, we have the same problem. This time it is for service improvement rather than sanctioning, but both are done by the same organisations. LCR therefore thinks this should not be done within UWV or SVB themselves, but for instance at Inlichtingenbureau or BKWI. These organisations are responsible for data coupling and you could make special legislation for this, and cumulate learning effects on how to build and supervise algorithms in the setting of administrative data coupling. By putting this at external organisations, it is also for external supervision, such as AP, where to look.

What is the contact between citizens and the LCR like?

Some citizens we look for proactively, some come to us. People come to us with issues and we also have campaigns on websites, people respond to those, send in information or have questions. We try to help people and talk to them. When we come across a case, we have a sort of triage model to decide which cases we look at more intensively. Those we can handle in

depth and see what people went through, if there is a policy problem, and we investigate this and publish solicited and unsolicited advice. Apart from this we have membership organisations and those have members themselves. Those can also raise problems. If there are bigger problems, those cases are also brought to us.

But you are not that involved when it comes to algorithms being put into use?

Only when there are issues or if we recognize that there is an algorithm involved. But we are not often proactively involved.

Should LCR be involved proactively in development phases of model development?

As I said, not everyone can oversee the implications and societal effects of certain models. I raise my doubts when I see such things in research guiding committees LCR is involved in, but we are not always at the table. The client council of UWV is sometimes involved when there are issues, they then hire someone that can help them look at an algorithm so they can advise the board. It is always reactive, but as a governmental organisation it would be good to proactively do this. That step has not been made yet, but is something organisations have to get used to and discover the added value of. It would be good to organise such a thing centrally as a government, which is why we are proponents of AP doing this instead of it happening within executing agencies. There are a lot of cases that LCR can impossibly all be involved in. But the government could ensure the weighing of interests is centralised, involving employers, clients, patients and patient organisations, people with disabilities, and all sorts of interest groups, some of which are related to LCR. If you only chase incidents, then the learning effects and do's and don'ts are never centrally organised. If we are part of a guiding committee and state our position, then a certain decision is made and the committee is dissolved, but our input and the knowledge created in these committees is not logged. If this is done, and it is documented, this would be educational, also when assessing other algorithms. AP as a central player could organise these interests and try to capture learning effects, where it would be interesting to ensure that interest groups have a seat at the table. There currently is no such place. I think it is time to think about such a thing. It would be important to do this several times a year. LCR sees a lot of policy propositions, and we could use the learning effects to properly make assessments. Now it depends on whether there are people with certain knowledge present, but you want to safeguard it institutionally and make sure it is durable. Then you can invest in such knowledge.

So that is about proactively and reactively taking action and maintaining knowledge?

Yes, because when legislation passes by us, I sometimes wonder if things are properly safeguarded, for instance when it comes to privacy law. The target group register (*doelgroepregister)* is a good example, which is tied to the jobs agreement (*banenafspraak)*. People with a disability are filed in this register, with the goal of ensuring extra jobs were created for those with disabilities. Employers that didn't adhere to the jobs agreement could be sanctioned. We are more than six years in, and sanctioning does not happen. So I wonder if it a good thing, also from a privacy perspective, as many people are filed in the register.

So that also ties to the issue of drawing out consequences and meaning that you mentioned?

Yes: what are the consequences? Try to logically reason what the consequences will be. But when LCR criticizes such things, we do not always get invited anymore. There is still a

discussion, but it is about whether the register should be made more accessible. Which is the wrong discussion. Things are seen from a certain political reality, where goals are meant to be reached and certain agreements have been made. But I wonder if it reaches the effects it is meant to reach. You see these things happening in real life, and if we were to organise it from scratch tomorrow, I wonder if we would do it in the same way.

We talked about enforcement and proactive service, there are also systems working to automatically allocate benefits. What do you see happening there?

When it comes to the assumptions, we see that the reality of citizens is more complex than what is put into the models. When it came to Wajong, example calculations were made to see how much benefits citizens would receive. A model is made for this purpose, in order to organise a standard process, but it doesn't keep into account exceptions. The reality of the benefits is more complex than the models can account for, because the models are organised per law. But then we see that some citizens get a bonus, or a thirteenth month pay, or they are paid every four weeks instead of monthly, or they only work a fraction of the time. The models do not account for all these other things, leading to disappointment over the calculations. A model give a number, when I then look at it, I get a different number. And when we ask for a recalculation, the executing organisation gets a third number. Who can tell me which is right, also with the risk of having to pay it back in mind? The complexity of the reality people face means models are made that can only attend for 80% of the requests. But the remaining 20% is so complex, and they are put through the same standard process, which leads to frustration. Especially when they have to pay money back after a certain amount of time.

What does not help is that we see that the goal is currently to be precise to the single euro, and people experience income insecurity because of it. And it also effects other benefits, because your income benefit [uitkering] influences other benefits. So first, people try to properly calculate their income, and then they see the effects on their rent subsidies, healthcare benefits, and so on. The interdependencies we built into the system when it comes to income are not traceable anymore for the average citizen. And that means the normal process leads to a lot of problems.

A solution would be to adhere to different tolerances in this system for when these interrelated benefits are changed. If there is a problem, currently the bandwidth within which it is enforced or recalculated is a single euro. The idea could also be: we understand the complexity of the law, so we adhere to broader margins, and if it is within those margins, it is our mistake and we will not recalculate and reclaim. That possibly leads to fewer people faced with having to pay back benefits. When we build in currently non-existing, broader tolerances when it comes to income, this also influences other benefits. We should always keep into account that the true system is more complex than we can organise with IT. But we seem to have lost that way of reasoning, because: the computer never makes a mistake, right?

D.8 Policy Advisor AI and Algorithms at Ministry of the Interior (I8)

What does your work regarding algorithmic systems entail?

Respondent is part of a team that deals with digital society, which means the digitalisation of society and resulting need for government action. Respondent specifically works on AI and Algorithms. The team works along several axes:

National/international: part of the work relates to law and policy for AI and algorithms for the Netherlands, but other things also concern the rest of the world. The team works on the implementation and execution of the AI Act, which part of the team helped negotiate. Furthermore, the team works on AI norms and guidelines within several international cooperations (e.g. UNESCO, UN, OECD).

Regulation/stimulation: another axis relates to, on one hand, a plethora of actions to regulate algorithms and AI, such as the algorithm register, an algorithm guideline (algoritmekader), and some tasks relating to the Directorate Coordination Algorithms (AP), such as agreements with the Minister (digitalisation), financing, and vision. On the other hand, BZK works on stimulating responsible AI in government. It is the lead of the workgroup public services of the Dutch AI coalition. The team looks at how public bodies shape use of AI, and works together with science, businesses, government, and civic society (citizens).

Only government/broader: a third axis relates to the policy work the team does on algorithms and AI in government. At the same time, the Minister is responsible for coordination of digitalisation policy, and thus the team is also involved in processes that are broader than just government, such as the Generative AI vision (visie op generatieve AI).

Furthermore, the team works on several cases involving algorithms (e.g. algorithmic decision making and the senate). All questions and requests from, for instance, the house of representatives are dealt with by the team respondent is part of.

That's quite a lot of work for one team!

Yes. We have a small team and around it are several projects we delegate to executing parties, people working on projects such as the algorithm register and algorithm guideline, and members of the national IT guild (rijks ICT gilde) who know a lot about data science and ethics. And a lot of work that is done, is executing motions and promises, movements within society, sometimes pushed for by DCA or AP. And there are certain requests that have been made part of the agenda value-driven digitalisation. One of these promises is a concept plan of how a government agency could shape governance surrounding AI and algorithms. So, governance specifically is a topic that is worked on.

When it comes to this concept and the algorithm guideline, they are quite overarching frameworks, how is everything brought together over a broad number and type of applications?

The goal is to bring everything together in one narrative and link things together. The guideline is not a traditional guideline, but more an overview of norms and requirements that are either mandatory by law of internationally seen as standards and that government should adhere to

when using AI and algorithms. The current process is putting the list of norms and requirements online and then, with the input of government and science, moving forward to trying to find common interpretations and then work towards practical guides. This can range

from checklists and forms to technical solutions to check for unwanted effects such as bias. The guideline is more and more the core of this work. The requirements stemming from the AI Act will, for instance, also be included in the guideline website. We also work together with supervisory authorities regarding the guideline to see if we can agree on what organisations need to comply with and how they can show they comply with these elements. Its not easy bringing this all together, but things get clearer along the way.

The guideline has a website that's quite open, you can give input. Are there certain signals that surface based on this process?

The guideline as once promised by the Minister was an overview of norms and requirements, but government agencies wondered what they needed to adhere to and how they could show this. This is partly a legal question, there are norms and requirements that apply either because they are legally binding or because they are generally accepted. There are various ethical frameworks that might not be legally binding, but consist of similar, generally accepted, elements (e.g. also in the framework by the Court of Audit). But, we are not just looking for such requirements, because simply saying that an organisation needs to adhere to something is not always useful. They for instance need to know: how do I know if my algorithm discriminates? What does it mean if my algorithm discriminates? There is no general consensus on this. The guideline is an attempt to reach consensus on explaining in simple terms what certain risks mean and what measures you could take to mitigate risks. We want everyone to be able to contribute and eventually the goal is to be able to agree on how things should be, broadly. There is this balance; some things may legally be the way they are, but you still need to agree on what this means in practice. That's what we try to do with this guideline.

Does this relate closely to the concept you mentioned?

The guideline originally was just a list. One of the other promises made by the minister to the house of representatives was to show how an organisation can organise their governance over AI and Algorithms. We think this logically relates to the guideline. By coupling the concept as a project to the guideline, things are brought together. For instance, we are working on the supervisory components of the AI Act. But at the same time, all the requirements from the Act need to become part of the overview in the guideline, and then government agencies will ask what it means for them. They then need practical guidance on shaping governance. If the act states systems cannot be used without meeting certain requirements, then organisations have to know how you can find these systems and get an overview.

When looking at bigger organisations such as UWV and SVB, this might be doable, but when looking at smaller organisations such as municipalities, isn't this tricky?

Bigger organisations generally have more instruments, such as ethical experts, data scientists, bias testers and so on. This means there are opportunities within these organisations. At the same time, that doesn't mean big organisations always have things in order and vice versa, smaller organisations can have a good grip on things. It's difficult to determine how things will be for

smaller organisations, but that is also the role of BZK: how can we help everyone with awareness on the topic? BZK is working on the governance aspects of this together with external advisors. Small organisations don't just have to deal with governance of algorithms

and AI, but they have to adhere to everything (datamanagement, cybersecurity, etc.), which is difficult as a small party.

And that's also partly why you want to bring everything together?

Yes, we want to lead the way and end up with something that is useful for everyone. For instance, in the algorithm register, municipalities that use similar software can use a template that was created for this software, adjust it, and thereby easily publish their algorithm. Perhaps a similar idea can be of use for governance of algorithms.

About the register: for whom is that at the moment?

Logically speaking for citizens, but they might not be directly interested. What we see in practice is that supervisory authorities including DCA keep an eye on the register, as do journalists (mainly investigative journalists), NGO's and researchers. So there are many types of visitors. It's all about transparency, an ethical norm in the world of algorithms and AI, but also something that follows from the Dutch general administrative law act (AWB), it has to be clear when algorithmic decision making is used. There are thus many visitors, some critical, which is important. With many different target audiences it is difficult to find descriptions that everyone can understand but at the same time serve a more expert audience.

A choice was made to keep it one register and not split it up, one for public and another for other parties?

It's one register. It would not be doable to make several registers. It is imaginable that when higher standards will be set by supervisors, politics or the science world, that this gets a spot in the current website. It is possible to add information, for instance in tabs, and serve different audiences. Those who want to know more just have to search a bit further.

Appendix E: Analysis of interviews

Table E 1 shows the different subthemes that were found as part of three major themes, together with the interviews these subthemes have been mentioned in. Subthemes were only included if mentioned in two or more interviews.

Table E 1 Analysis of interviews

(Sub)themes	**Interviews**							
Observations on governance of AS	**I1**	**I2**	**I3**	**I4**	**I5**	**I6**	**I7**	**I8**
Average citizen not aware of use of AS	x					x	x	
Interdependencies and data coupling not traceable for citizens	x						x	
Bottom-up signalling by citizens of AS involvement	x					x	x	
Supervisory organisations do not always have direct knowledge of AS involvement	x			x		x	x	
Knowledge not well documented and retained	x				x		x	
Cases provide tangible knowledge	x				x	x	x	
Organisations have difficulty translating values and human rights to AS in use		x			x	x		x
Gap between technical teams and legal, ethical, political issues					x	x		
Good technical monitoring, but difficulty accounting for all relevant factors					x		x	
Algorithm register is incomplete and not attuned to citizens' needs		x	x	x		x		
Plethora of (overlapping) laws, regulations, frameworks, guidelines			x		x			x
Existing tools do not guarantee safety			x		x		x	
Organisations have awareness and want to do things right			x	x	x		x	
Sufficient measures are not defined, causing insecurity			x	x	x			x
Attention for AS and outrage when mistakes happen			x		x	x		
Governance improvements in general landscape	**I1**	**I2**	**I3**	**I4**	**I5**	**I6**	**I7**	**I8**
Central and proactive supervision	x	x		x		x	x	
Independent supervision on risk scans if not disclosed		x				x		
Knowledge development and retention	x		x		x	x	x	
(Further) specification of sufficient measures and explanation of human rights		x	x		x	x		x
Different political and societal discourses			x		x			
Concrete cases and jurisprudence to learn from	x				x			
More complete algorithm register		x		x		x		
Governance improvements within SUWI	**I1**	**I2**	**I3**	**I4**	**I5**	**I6**	**I7**	**I8**
Increased transparency, motivation of decisions and actions	x	x		x		x		x
Centralisation of certain tasks	x						x	
(Proactive) connection to citizens and societal groups	x				x		x	
Culture and multidisciplinary discussion					x		x	
Integral trade-offs on use of AS				x	x		x	
Application of existing and upcoming laws and regulations		x		x	x	x		x