# Controlling the Point of Emergence

## Safety control interventions to curb digital cages in social welfare

M.F. Enbergs

# Controlling the Point of Emergence

## Safety control interventions to curb digital cages in social welfare

by

M.F. Enbergs

to obtain the degree of Master of Science
at the Delft University of Technology,
to be defended publicly on Wednesday May 22nd, 2024 at 11:30 AM.

| | | |
|---|---|---|
| Student number: | 5635209 | |
| Project duration: | November 27th, 2023 – May 22nd, 2024 | |
| Thesis committee: | Dr.ir., R.I.J. Dobbe, | TU Delft, supervisor |
| | Dr., B. Wagner, | TU Delft, second supervisor |
| | M.Sc. PhD candidate, S.J.J. Nouws, | TU Delft, |
| | Prof. Dr., N. Bharosa, | TU Delft, chair |

TU Delft

# Preface

*M.F. Enbergs*
*Delft, April 2024*

Before you lies the master thesis titled: ”Controlling the Point of Emergence Safety control interventions to curb digital cages in social welfare”. This thesis was written to obtain the degree of master in the study of ”Complex Systems Engineering and Management” from October 2021 to May 2024.

The motivation for this thesis stems from a number of factors that I will outline in the following paragraphs.

During my studies, I frequently saw myself drawn to the topic of artificial intelligence. My motivation stemmed from the understanding that artificial intelligence, as a disruptive innovation, harnesses great potential to improve our future lives. During my studies, I also followed the public debates held about AI in the media that are primarily preoccupied with bouncing scenarios of utopian salvation and dystopian destruction back and forth. While scientific discussions about AI provide a more differentiated perspective they, in my view, also tend to provide an isolated perspective on the technology. While such a perspective undoubtedly has its merits, this inadvertent reductionism also leads to a loss of information. Through various lectures and assignments and the nature of studies at TPM faculty, I began to develop a more nuanced perspective on this issue and shifted from a strong technocratic view to a socio-technical perspective. I began to view artificial intelligence as a subpart in a larger web of interdependent components. I began to realize that AI though it has great positive potential, can only be realized as a force for good in conjunction with its surrounding socio-technical environment.

As with any rapidly developing technology our understanding of it is still limited. This leads us to overlook the complexity of interrelations between the technology and its surrounding environment. Which exposes us to greater risks because it enhances the curse of flexibility. Systems, such as AI systems, are no longer limited by physical or materialistic constraints, while this increases capabilities it also results in the removal of various, ’natural’ safety barriers. As these systems become increasingly powerful and influential, they inadvertently impact our lives on more fundamental levels. Problematic in this regard is that we tend to single out specific system components as causes for an error without realizing that harmful outcomes can only be created in conjunction with the environment that surrounds these components. This leads me to the understanding that AI though it has great power, also possesses weaknesses. The technology needs to be managed and *will only be effective if deployed within the right context*. This is important to understand to manage technology for good, yet often overlooked. Within our excitement for technology we sometimes purely focus on the positives and forget or disregard the duality of technology. Our capacity to focus on the inherently positive things we can in-vision, undoubtedly a good human trade, leads us to overestimate their likelihood of occurrence. This fundamental reflex of our human physiology sometimes tricks us into creating SISPs (Solution in search of a Problem). We create solutions for problems that, in this specific form, only exist in our minds. In our excitement, we forget to pause and verify if the assumptions we make about the technology we use and the environment that surrounds us are correct. The result?

We underestimate the potential for harm that technology can exert. We forget that technology for all is liberating characteristics also ensnares us to its natural characteristics and rationality.

Moreover, I sought to understand this issue from a practical perspective. As mentioned previously AI only really ”comes to life” in the interplay with its environmental context. In research, we often focus on the inherent features of a technology or system, looking for ways to enhance the technology itself. However, while such focus has its merit, it do not recognize that outcomes are generated while a system is operated in the real world rather than analyzed in a design vacuum inadvertently filled with abstractions, reductions, and assumptions. We fail to recognize that its true characteristics emerge only when deployed in a real-world environment. This reductionist view leads us to lose the behavior characteristics that a system begins to exert based upon the interactions between its technological, social, or institutional components. The unpredictable interactions within such environments are crucial, and potentially even more important, for the successful application of technologies like AI. Which ultimately lessens our understanding of how to manage and utilize technology, AI, for good.

While there are always things one can regret about one's choice to study a particular issue or subject, it is important to pause and consider the perspectives and knowledge one has gained. In my case, I am grateful for the deeper understanding of the interplay between technology and society I was able to obtain through my choice of study.

To finalize my thoughts I want to provide the reader with the following quote, which provides ”fruit for thought” if applied to the topic discussed in this thesis:

> ”The caterpillar is a prisoner to the streets that conceived it.
> Its only job is to eat or consume everything around it,
> in order to protect itself from this mad city.
> While consuming its environment the Caterpillar begins to notice ways to survive.
> One thing it noticed is how much the World shuns him, but praises the butterfly.
> The butterfly represents the talent, the thoughtfulness, and the beauty within the caterpillar,
> but having a harsh outlook on life the caterpillar sees the Butterfly as weak
> and figures out a way to pimp it to his own benefits.
> Already surrounded by this mad city the caterpillar goes to work on the cocoon which institutionalizes him.
> He can no longer see past his own thoughts.
> He's trapped.
> When trapped inside these walls certain ideas take roots,
> such as going home, and bringing back new concepts to this mad city.
> The result?
> Wings begin to emerge, breaking the cycle of feeling stagnant
> Finally free, the butterfly sheds light on situations that the caterpillar never considered, ending the internal struggle.
> Although the butterfly and caterpillar are completely different,
> they are one and the same.”

– Kendrick Lamar, Mortal Man
(Lamar et al., 2015)

# Executive Summary

The thesis explores the complexities and risks associated with Automated Decision-Making (ADM) systems in social welfare, focusing on the phenomenon of ”digital cages.” The digital cage concept refers to a situation where rigid systems, perpetuated through algorithmic misclassification, inadvertently trap citizens in bureaucratic complications without recourse. This thesis analyzes the digital cages concept from a system safety perspective, on the example of the Dutch Toeslagenaffair. The thesis centers on understanding how digital cages form within social welfare systems and seeks methods to mitigate their emergence through targeted interventions using system safety theory. The pivotal question addressed is:

***”What are safety control interventions to curb the emergence of algorithmic decision-making systems-induced digital cages in the context of national social welfare administration in the Netherlands?”***

The research uses design science to derive control interventions that help curb the emergence of digital cages in social welfare. To this purpose, it employs a combination of methodologies including System Theoretic Process Analysis (STPA), literature reviews, and an evaluation workshop to examine the causes and consequences of failures in ADM systems. This comprehensive approach helps identify specific points where interventions can be strategically implemented to prevent the onset of digital cages.

In a first step, the research delineates the issue of digital cages from a system thinking perspective. A particular focus is set on the phenomenon of emergence. From this analysis, the importance of a holistic system safety approach is deducted.

Hereinafter the thesis moves toward the unit of analysis by conducting a socio-technical analysis of the problem space present during the time of toeslagenaffair. This analysis is conducted in order to inform the subsequent safety ’Systems Theoretic Process Analysis’ STPA. STPA is utilized to identify system hazards present in the Belastingdienst ”Toeslagen” system at the time. These hazards are derived by analyzing the present hierarchical control structure for unsafe control actions. Ultimately, the system hazards build the biases from which system constraints can be derived. From this analysis, a broad inability to identify, control, and recover system states was detected in the system.

Subsequently, the thesis moved towards deriving system safety constraints that serve as design requirements for the subsequent intervention design. In systems, safety theory plays a pivotal role in order to restrain a system’s actions to safe behavior. System constraints built a natural extension to safety hazards. Within this research 6 main system constraints with several sub-constraints each were identified. These constraints are directed towards enabling a safe system operation by addressing the problems previously identified in hazard analysis.

Thirdly, the research moves on to design the control interventions that help to curb digital cages in social welfare. The thesis proposes assumption-based leading indicators entailed in a program as the main intervention, which serves as proactive measures in predicting and mitigating potential failures before they manifest in the operational ADM system. These assumption-based leading indicators are derived from the foundational assumptions about how the ADM systems are expected to behave and interact within the broader bureaucratic system landscape they are implemented. Assumption-based leading indicators offer a proactive approach to safety management. By monitoring these indicators, organizations can identify when a system is approaching its safety constraints and take corrective action before a constraint is violated, thereby preventing accidents or system failures. As leading indicators provide early warnings of potential safety issues, they can inform the operators when there is an increased risk that safety constraints may be violated. They also point towards the need to revisit

and possibly revise safety constraints to better reflect the current understanding of the system and its environment, since once an assumption has ”failed” the integrity of the underlying safety design of a system is threatened. There are three ways in which lead indicators can be enforced:

- Shaping actions to prevent violation of the assumptions
- Hedging actions to prepare for the failure of an assumption
- Assumption checking during operations

  with Signposts to trigger specific checks

  Checking during system operation (periodic or continual)

Several sub-interventions are proposed to aid in the enforcement of the leading indicators. They are diverse in nature and entail both technical as well as organizational solutions.

Finally, the designed intervention is evaluated with the help of a workshop. The conducted evaluation points towards the potential of organizational interventions for safety improvement. Additionally, it also highlights the need for a more holistic problem approach that enables stakeholders to combat the issue of digital cages in an interdisciplinary fashion. Moreover emphasizes the need for further evaluation, specifically the need for empirical studies that can provide quantitative proof regarding the utility of the proposed interventions.

The research concludes by highlighting the potential of assumption-based leading indicators. Assumption-based leading indicators are predictive measures that rely on underlying assumptions about how certain inputs or actions correlate with future outcomes. The leading indicators provide a novel approach to enhancing system safety in social welfare administration by addressing the underlying assumptions of the system in operation. This offers a substantial potential to transform how social welfare systems manage and utilize ADM, potentially reducing the incidence of digital cages. By detecting early signs of hazardous systems states digital cages can be mitigated before they solidify into systemic issues.

The thesis not only addresses the immediate concerns regarding digital cages but also opens pathways for future inquiries into safer and more equitable administrative practices. Future research is suggested to further refine these indicators and explore their applicability in diverse administrative contexts, ensuring they can effectively adapt to the evolving landscape of public administration and technology.

# Contents

# List of Figures

# List of Tables

# 1 Introduction

This chapter serves as an introduction to the thesis topic by first providing context to the research area. Section 1.1 summarizes the trend towards increased utilization of automated decision-making systems in public administration (see section 1.1). Secondly, the introduction goes on to explain why these developments matter and hence why digital cage research is important (section 1.2). Thirdly, section 1.3 continues the introduction by outlining the main research area by giving an explanation of the digital cage phenomenon and relevant conceptualizations surrounding the issue. It exemplifies how complex systems thinking (section 1.3.2) as well as the system safety methodology (section 1.3.3) may help to understand and alleviate the issue. Finally, section 1.4 outlines the structure of the subsequent thesis and gives a brief overview of the different parties involved.

## 1.1. The use of ADM systems in modern Government

In recent decades we have witnessed an increased push towards automation in an effort to improve efficiency, scalability, and profitability. In the age of globalization and interconnectedness humanity's relentless pursuit of productivity gains is fed through constant innovation and development of new technologies. The rise of artificial intelligence as well as continuous improvement of machine learning techniques are two of the underlying technological innovations driving this spiral (Oosthuizen, 2022). These evolutionary forces can also be witnessed in the public sector. Encouraged by the promise that AI will radically alter governance for the better, make it ’smarter’ (Gordon, 2004), officials have been busy launching a number of digitization offensives aimed towards automating and augmenting public decisions.

Yet, contrary to mainstream expectations, many of the promised positive ’transformative effects’ (Moran et al., 2006) of algorithmic systems have not materialized, while inherent risks of these systems have aided in the emergence of new forms of harms affecting citizens (Veale et al., 2018). In multiple public cases the use of AI in the form of automated decision-making (ADM) systems has led to a centralization of decision-making resulting in the obstruction of street-level discretion, see (Peeters & Widlak, 2018) also (Bovens & Zouridis, 2002). This centralization of decision-making represents a ’fundamental change of character’ of public agencies converting the internal organizational structures from traditional street-level bureaucracy to system-level (Bovens & Zouridis, 2002) to infrastructure-level bureaucracies (Peeters & Widlak, 2023), and has far-reaching implications on how modern government works and consequently interacts with their constituents. A large number of the deployed digitization projects saw public policymakers augment ’primary processes’ of governance with systems that lacked appreciation for the nuances and complexity of public governance. Peeters and Widlak point out, that the algorithmic systems, at the center of these projects, often wrongly classified citizens, placing new forms of administrative burdens unto them that resemble the characteristics of Weber´s famous ’iron cage’ (Weber, 2016) see also (Peeters & Widlak, 2018). The frequently questionable decisions made or supported by such systems led to multiple government scandals in a number of different nations. Prominent examples, widely discussed in the news, in which ADM systems falsely classified citizens, are for example the Dutch ”toeslagenaffaire” or the Australian ”Robodebt scandal”. These cases saw thousands of citizens falsely accused of social benefits fraud and forced to pay back large sums of money (Graycar & Masters, 2022) also (Heikkilä, 2022). The debt piled onto citizens resulted in the formation of a bureaucratic ”cage” that made it difficult or impossible for victims to find a job, receive other forms of benefits, or provide for their families (Henriques-Gomes, 2023). In some cases, the pressure of the debt led to parents losing their children to the authorities as their family fell below the minimum standard of living (Times, 2021). In other cases, the physiological toll of feeling ”trapped” and ”powerless” led some victims to commit suicide as they saw no way out of their situation (Mao, 2023). Much like in the famous 'iron cage' individuals subject to these cases have increasingly been trapped in expanding bureaucratic structures that reduced them to cogs in a machine (Peeters & Widlak, 2018) or rather data points in a computer. In contrast to Weber´s 'iron cage' in which bureaucratic rules and procedures stand at the center of the accused, these newly emergent structures share the distinct characteristic that an algorithmic entity represents the source of harm. Weber's 'iron cage' has transformed into a 'digital cage' (Peeters & Widlak, 2018).

## 1.2. Why it Matters

As illustrated above digital cages have the potential to inflict innocent citizens with severe harms, in many cases leaving them with irreparable damage to their 'welfare' and that of their families. It is clear that the ADM systems responsible, fail to meet the doctrine of the modern ideal of 'good governance' (Peeters & Widlak, 2023) and hence question its feasibility by threatening citizens and, inadvertently, the public's trust in government institutions.

In 2019 the UN's special rapporteur Philip Alston addressed the emerging consequences of a 'digital welfare state' in his ”Report of the Special rapporteur on extreme poverty and human rights” (Alston, 2019). Many of the negative externalities Alston outlines can be directly attributed to the emergency of digital cages. Namely, he expresses deep concern over the abundant evidence that already existing algorithmic decision support systems (ADM) are aiding the welfare state to predominantly target the poor and other marginalized groups of society while the well-off are not subject to equal scrutiny. He argues that with our current understanding of the digital welfare state, this trend will only be amplified ultimately resulting in a 'welfare dystopia' that will deny basic human rights and strengthen the increasing gap between rich and poor (Alston, 2019). As such digital cages pose a direct threat to some of the sustainable development targets put forth by the UN. Especially the goal to ”Reduce inequality within and among countries” (Assembly, 2015). Consequently, researching the fundamental factors underlying the emergence of 'digital cages' as well as mechanisms and interventions that help to prevent them represents a pressing matter of research. Based on the presented evidence, this especially applies to ADM systems deployed in social welfare administration, for reasons previously eluded to by Alston.

The social welfare system exhibits complex characteristics that precipitate the emergence of the digital cage, these characteristics are (Brazier et al., 2018):

- it is a system of systems
- it has a large number of (emerging) uncertainties
- it involves many stakeholders and interdependencies
- it is dynamic and under continuous change
- the structure itself exhibits emergence, openness, and learning

Designing in the context of such a system requires a multi-disciplinary approach that is conscious of the peculiarities of complex systems and adapts its design to these constraints. A complex systems engineering approach with particular emphasis on socio-technical systems therefore represents an excellent point of departure for this research project. Conversely, the subject fits well into the profile of research conducted at the TPM faculty of TU Delft.

Subsequently, this thesis will delve deeper into the scaffold of the digital cage and its connection to complex systems. The next chapter will lay out why the structure of the surrounding system represents a key component to the formation of digital cages.

## 1.3. Introduction to the Digital cage

### 1.3.1. Scaffolds of a Cage

The ’digital cage’ is a phenomenon described by Peeters and Widlak as: ”an exclusionary infrastructure that hinges on information architecture instead of Weberian rules and procedures” (Peeters & Widlak, 2018). Nouws et al. characterizes digital cages as action situations where the: ”rigidity of algorithms and information architectures [...] results in automation creating its own reality” (Nouws et al., 2022). Digital cages hence create a situation in which bureaucratic ’red tape’ (Bozeman, 1993) and catch-22´s facilitated through algorithmic classification and prediction tools, contribute to unforeseen hazards in terms of ‘administrative exclusion’ (Brodkin & Majmundar, 2010) and discrimination (Peeters & Widlak, 2018) against citizens. Inadvertently, based on these hazardous systems states, harms emerge. Often in the form of ’ripple effects’ that propagate through the different levels of the bureaucratic organizations in many cases even transcending onto different branches of governance (Pel, 2022). Ultimately engulfing the citizens, subject to bureaucratic decision-making, in ’cages’ they are not able to escape from. The resulting exclusion hence is a consequence of citizens’ powerlessness to overcome administrative burdens (Peeters & Widlak, 2023). The harms such a cage may project onto its subjects vary in their form. Victims often see their privacy violated, the ’burden of proof’ amended to the benefit of the state, are subject to disciplinary decisions whose effects propagate through multiple facets of their life, experience a clear lack of support infrastructure, are perceived by the system to be the perpetrator of wrongful action and often suffer severe financial consequences detrimental to their standard of living. The ’digital cage’ hence represents a subsidiary of the ’iron cage’ as a concept for the predicament of modern human beings trapped in the socioeconomic structure they help to create (Baehr, 2001). In contrast to its origin term the ’digital cage’ however exhibits stronger characteristics of exclusion not merely forming a “shell as hard as steel”, confining individuals (Baehr, 2001) and imposing restrictions on their freedom, but rather inflicting direct, harmful acts of intervention on them. Additionally, another distinction is that in the ’digital cage’ civil servants no longer hold authority through exercising discretion. Street-level bureaucrats themselves are confined to work within the rationality of the information systems made for them by software developers and IT engineers (Peeters & Widlak, 2018). Consequently, as Peeters and Widlak point out they: ”[...] are incapable of making value decisions” (Peeters & Widlak, 2018).

Researchers therefore have exclaimed these systems to be the work of ’administative evil’ (Graycar & Masters, 2022) perpetrated by high-level policy makers. Yet, as the previous explanations show such characterizations fall short of capturing the complexity involved in the formation of a digital cage. It rather seems that the information architecture, and the infrastructure it is embedded in, exhibits complex characteristics that form the structure for digital cages to ’emerge’ as a systemic phenomenon (Nouws et al., 2022). Consequently, just as Weber’s ’iron cage’ emerged as an ”order [...] now bound to the technical and economic conditions of machine production which [...] determine the lives of all the individuals who are born into this mechanism” (Baehr, 2001), so the ’digital cage’ emerged from the nature of the infrastructure it was conceived to and restraints actors to this structure. Ultimately, however, both the ’iron-’ and ’digital cage’ only materialize in the interplay with humans. The structure itself only becomes important through the effects and constraints it imposes on individuals. It carries no purpose by itself. Due to this interplay, the scaffolds of the ’digital cage’ can only be fully described from a socio-technical perspective, that captures the impact the system structure has on humans subject to it. The following section utilizes a systems perspective and its related concepts to further describe the phenomenon of the ’digital cage’ within the social welfare system.

### 1.3.2. Social Welfare: Digital Cages emerge in a complex System of Systems

Through analysis of the digital cage in section 1.3.1 it becomes apparent that the formation of digital cages is a phenomenon of a systemic nature. Meaning, the structure of the system itself leads to the creation of the digital cage. Complex systems thinking and its concepts can help to make sense of such a development. Complex systems analyze a system from a holistic perspective that can provide insight into peculiar behavior the systems at hand may exhibit. One of these peculiarities especially important to the concept of the digital cage phenomenon is ’emergence’. ’Emergence’ in system literature refers to: ”a process of interaction between lower-level system properties that creates non-obvious consequences over time” (Corning, 2015).

Figure 1.1 illustrates this process. The emergence of a digital cage viewed from a system think-

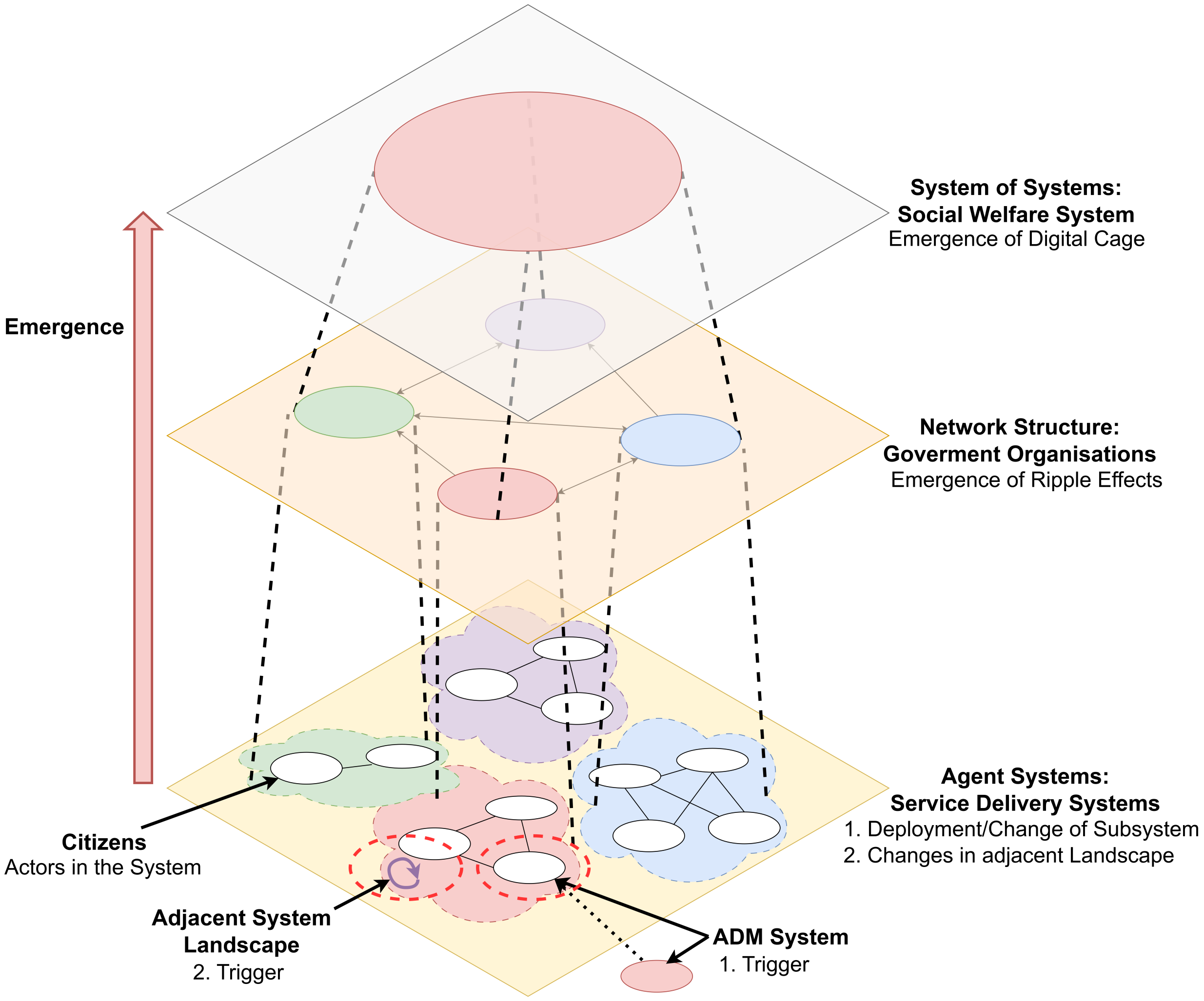


Figure 1.1: ADM in the Welfare System of Systems

ing perspective is a process triggered through complex interactions between different subsystems that follow diverging goals and objectives. The emergent phenomenon, the ”digital cage”, can only be observed at a higher system level, once its effects are impacting individuals. The cause however lies in the interaction between lower-level system components or subsystems of the social welfare system. Emergence is a phenomenon especially present in such complex systems because complex systems are dynamic networks consisting of many agents and subsystems operating in parallel and thereby constantly acting and reacting to what other agents or subsystems are doing (Waldrop, 1993). The control of a system tends to be highly dispersed and decentralized (Waldrop, 1993) as visible in the social welfare system. Consequently, the overall behavior of the system is the result of a huge number of decisions made every moment by many individual agents and subsystems (Waldrop, 1993). Such effects are exacerbated in a system with elaborate vertical differentiation (Allen et al., 1999). In such a system, changing the system typology by implementing a new subsystem or changing an existing one can result in a causal chain of behavioral adaption that is non-obvious and difficult to oversee. Such triggered ’ripple effects’ ultimately propagate through the system and can sum up to form a digital cage. Applied to the context of ADM systems in social welfare this means digital cages can emerge in two ways, see Figure 1.1:

1. Through **direct** changes to the ADM system. Either by making changes in existing ADM system or through implementing an entirely new system. Or through inherent mutation of the ADM system. Such mutation affects ADM systems over time, and so are machine learning applications in production often subject to data and concept drift (Côté et al., 2023), see also (Sculley et al., 2015).

2. Through **indirect** changes. Occurring in the system landscape surrounding the ADM system. These changes can occur through deliberate action by changing or implementing an adjacent subsystem, but they can also be unintended changes, due to evolving processes and practices. Ultimately, these indirect changes can alter the interaction patterns of the ADM system with its surrounding system landscape, diverging the system away from its indented process structure, and inadvertently enabling the emergence of a digital cage.

Given the ongoing trend of ever-increasing utilization and interconnection of IT and ADM systems in governance, this problem is likely to be exacerbated in the future. Since the two main catalysts for emergence, interconnection or interaction between subsystems, and a number of subsystems, are increasing. Previous research in this field has been focused on the characteristics of ’digital cages’, the causes for their emergence as well as on harms that arise through them (Peeters & Widlak, 2018) see also (Pel, 2022). Scholars have been analyzing the design methodology of public administration and identifying it as a key contributor to the emergence of digital cages. Crucially, however, there seems to be a heavy emphasis on the facilitating factors and improving the design practices of ADM systems, while few address issues in the area of ADM system implementation and -operation triggering these cages. Given the emergent nature of digital cages, hence their non-obviousness, this focus may prove to be presumptuous.

The focus on the ADM system’s early stage design practices overlooks that the systems are utilized in a complex system. Such systems exhibit chaotic characteristics, which make them extremely sensitive to system state conditions. Hence small system status discrepancies between the early-stage design environment of an ADM system and the deployment/operation environment can trigger unforeseen skewed and distorted outcomes at aggregated system levels. Consequently, while sound design practices in the initial state of an ADM project are necessary they do not infer that the system once deployed and operated in a larger system of systems environment will behave as intended. On the contrary, from system safety engineering we learn that the optimization of individual components or subsystems does not necessitate a general performance improvement of the overall system. In fact, improvement of a particular subsystem are likely to actually worsen the overall system performance because of complex, nonlinear interactions among the components (N. Leveson, 2011). Literature addressing complex systems engineering has shown that sound system deployment and operation methods are key to effective interventions in evolving socio-technical systems (Brazier et al., 2018). Consequently, the focus of this research will be set on ADM system deployment and operation in social welfare from a systems perspective.

A socio-technical analysis, within the context of the complex system perspective, involves examining the interplay between social and technical components within a system. It recognizes that technology and social factors are intertwined and that changes in one can affect the other. Such inferences are vitally important for the study of ’digital cages’ because emergent phenomenons can only be understood from a holistic perspective as ”the sum is greater than its parts” (Corning, 2015). This perspective allows us to view the digital cage as a result of structural dynamics, giving designers the ability to not only look at technical errors that may lead to an accident, but underlying causes. Such a system perspective provides a framework for understanding, managing, and designing systems that are characterized by intricate interactions and dependencies. Consequently, it can greatly help to improve decision-making, adaptability, and the ability to address challenges in a more holistic and effective manner. Each of which is highly important to tackle emergent problems.

### 1.3.3. The System Safety perspective

Little has been written about the crucial role of the ’point of emergence’ that the process of ADM systems deployment or operation in the social welfare system represents. However as previously eluded to, ensuring the safety of such systems is vital to the overall effectiveness of the principles of ’good governance’. One therefore particularly important perspective for the study of ’digital cages’ is system safety. The system safety approach is uniquely suitable for exploring the failure leading to ’digital cages’ as it provides a holistic perspective on system structures that allows researchers to identify dysfunctional interactions among system components (N. Leveson, 2011). Therefore, it is able to identify emerging behaviors within a system that were not anticipated during the design phase. Ultimately, enabling designers to come up with suitable mitigation techniques.

System Safety focuses on the identification, assessment, and management of risks associated with complex systems to ensure their safe operation (N. Leveson, 2011). It is therefore particularly crucial in situations where failure of a system can lead to catastrophic consequences, such as a digital cage.

The primary goal of System Safety is to prevent accidents and mitigate the impact of potential hazards or eliminate them from the safety control structure of the system. System safety recognizes that complex systems often migrate to a higher risk level (N. Leveson, 2011) also (Rasmussen, 1997). As Rasmussen points out this migration is often caused by 'adaptation' from actors under pressure to reach "cost-effectiveness" and "efficiency" (Rasmussen, 1997). According to him, however, 'adaptation' is a manageable process (Rasmussen, 1997). Therefore in order to manage the migration of the system moving to a higher risk level, one must manage the process of 'adaptation' (N. Leveson, 2011). 'Adaptation' is triggered through changes within the systems. For example the deployment of a new ADM system as well as adjustments performed on an existing ADM system or its surrounding systems landscape (e.g. other subsystems).

Therefore if the goal of this research is to make the process of ADM deployment and its subsequent operation safer, we: **"must consider the processes involved in accidents and not simply events and conditions"** (N. Leveson, 2011). Consequently, 'process control' of ADM deployment and operation must be carried out.

Process control in system safety is exhibited with the help of a *system safety control structure*. In the context of system safety, a control structure refers to the set of mechanisms, processes, and procedures designed to regulate, monitor, and manage the behavior of a system to ensure its safety and reliability (N. Leveson, 2011). One very important mechanism of safety control structures are 'control actions'. Control actions help to impose safety constraints on system behavior to ensure it is operated within a safe state (N. Leveson, 2011). Control actions can be passive, meaning they maintain safety by their presence, or active, meaning some action to provide protection is necessary (N. Leveson, 2011). Active control actions are comprised out of the detection of a hazardous event or condition, measurement of some variable(s), interpretation of the measurement (diagnosis), and response or recovery from a failed system state (N. Leveson, 2011). Each of these actions must be conducted to fully complete 'active control'. Due to the introduction of ever more sophisticated technologies, especially information technologies, the number of 'passive controls' within existing complex systems is decreasing. Passive controls are often based upon physical principles such as gravity, distance, and time (N. Leveson, 2011). Modern technologies help us overcome or bend these physical principals thereby inadvertently removing passive controls (N. Leveson, 2011). This effect further perpetuates the migration towards higher levels of risk within the system, because previously functional, passive control mechanisms no longer constrain the system. Simultaneously, this makes the need for 'active control' interventions within existing systems more apparent.

As a result, the importance of ADM system deployment and changes are increased, as it represents 'direct adaptation' of the system. Hence, if it is carried out with insufficient 'process control' it inadvertently perpetuates the systems migration to a higher level of risk. 'Adaptation' however goes beyond the mere deployment of a system, as it can occur at any time in the subsequent system operations. Control actions/interventions must therefore also help to ensure that the deployment process itself as well as any operational adaptation made are confounded within the boundaries of a safe system state. Hence it is the goal of this research to identify and illuminate 'safety control interventions' in ADM deployment and operations in social welfare. To give administrators instruments they can utilize to make ADM systems and, as a consequence, the social welfare system itself safer.

### 1.3.4. Control Actions/Interventions in Practice

In complex systems ensuring 'process control' through control actions/interventions is highly important to the overall system health. However subsequent analysis of digital cages that emerged in practice showed that such safety concepts were often not sufficiently introduced. For example in the Dutch child care benefit scandal or 'toeslagenaffaire', a subsequent report conducted by an independent committee found that the ADM system in question (FSV) was 'vervuild' (rotten), through use by the Belastingdienst (tax authorities) (Financiën, 2022). Additionally, the report points to the lack of decision tractability through missing logging function, lack of clear guidelines regarding the use of the ADM system, the

lack of control regarding access to the system both internally as well as externally, and the increased build-up of ’technical debt’ insufficient service of the system infrastructure (Financiën, 2022).
All the above-provided examples present clear safety hazards, yet from system safety theory we learn that the biggest fault committed in the affair is the clear lack of control actions/interventions. Unreliable systems can still be operated safely if hazards are kept in check by control actions/interventions that allow the system to ’fail into a safe state’ (N. Leveson, 2011). Hence the pivotal factor in the toeslagenaffair was the clear lack of monitoring and feedback mechanisms. With such mechanisms in place during deployment and operations hazardous system states could have been detected, measured, interpreted, and recovered.
Missing control actions/interventions for ADM systems during deployment and operations therefore represent a core problem in the emergence of digital cages and hence must be addressed. Ultimately, monitoring and feedback mechanisms are pivotal to the improvement of any system design. They also represent a prerequisite to any attempt at iterative design.

## 1.4. Thesis Structure

The previous chapters have provided an overview of the importance of researching the phenomenon of digital cages, the necessity to closely analyze the issue of deployment as the ’point of emergence’, and the clear need for a more safety-focused perspective on system adaptation, specifically for social welfare related ADM systems. The following chapters will subsequently analyze the issue of ADM system deployment and operation in social welfare. Starting with a literature review of the issue of digital cages, its causes, and possible objectives for solutions. Through this analysis, the need for deployment and operation-focused research will be explained further. Subsequently, the next chapter will introduce the main research question. It is followed by a methodology section that will define the research approach, the sub-research questions, as well as methods to be utilized during the research. Hereinafter the area of research will be analyzed through the introduction of the socio-technical environment surrounding the unit of analysis. This is followed by the identification of design requirements for the safety control interventions during ADM deployment in social welfare administration. These will then be formalized in Chapter 6.1. The research will be completed with an evaluation and a discussion of the designed interventions.

# 2

# Knowledge Gap & Research Question

The following literature review aims to provide an overview of the existing socio-technical factors precipitating the emergence of 'digital cages' as well as current design suggestions for mitigation to subsequently point out the need for research towards safety control interventions during ADM systems deployment and operations in social welfare. The chapter will then move to introduce the main research question.

## 2.1. Literature Review methodology

For this literature review the primary source was Scopus, while other search sources, drawn from throughout the research, were Google Scholar as well as "Connectedpapers.com". The initial search strings utilized were: "( TITLE-ABS-KEY ( adm AND systems ) AND ALL ( government ) AND TITLE-ABS-KEY ( public ) )", ( TITLE-ABS-KEY ( ai ) AND TITLE-ABS-KEY ( public AND administration ) ) and "( TITLE-ABS-KEY ( digital AND cage ) AND ALL ( government ) ) ". The results out of these search strings amounted up to 329 unique entities. Based on these results, a thorough analysis of the articles was conducted. Articles that did not sufficiently focus on automated decision making systems (ADM) in public administration were eliminated from the selection. Examples for such literature could be articles that addressed ADM systems in other contexts such as logistics of financial decision making or health care. Moreover, publications that did not substantiate further on the impact of ADM on digital cages and their emergence were excluded. Subsequently, the number of identified articles was boosted with the technique of backward snowballing in combination with "direct access sources" (sources known to the author). In total this yielded a number of 17 additional articles. The eligibility of the remaining articles was tested through gauging their respective consistence in regards to addressing ADM aided digital cages and their respective focus on the design and deployment methods in public administration. This literature search and selection process yielded 36 articles deemed suited for further analysis.

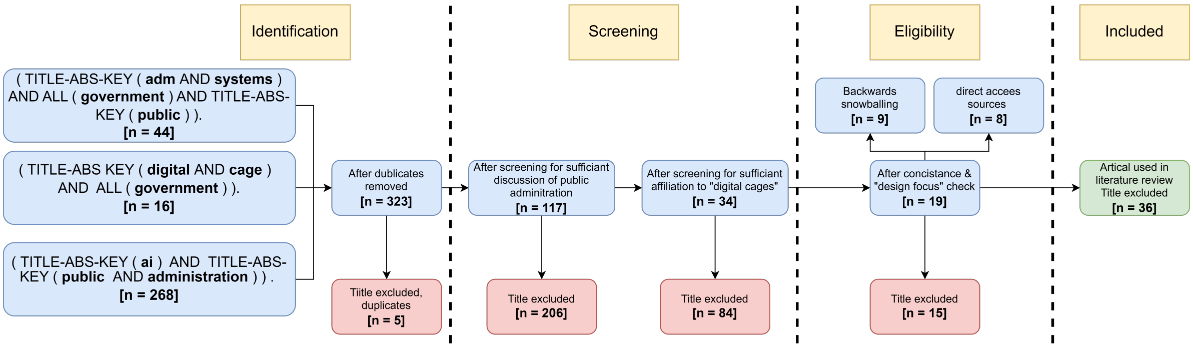

Identification
Screening
Eligibility
Included
( TITLE-ABS-KEY ( adm AND systems ) AND ALL ( government ) AND TITLE-ABS-KEY ( public ) ). [n = 44]
( TITLE-ABS KEY ( digital AND cage ) AND ALL ( government ) ). [n = 16]
( TITLE-ABS-KEY ( ai ) AND TITLE-ABS-KEY ( public AND administration ) ) . [n = 268]
After dublicates removed [n = 323]
Tiitle excluded, duplicates [n = 5]
After screening for sufficiant discussion of public adminitration [n = 117]
Title excluded [n = 206]
After screening for sufficiant affiliation to "digital cages" [n = 34]
Title excluded [n = 84]
Backwards snowballing [n = 9]
direct accees sources [n = 8]
After concistance & "design focus" check [n = 19]
Title excluded [n = 15]
Artical used in literature review Title excluded [n = 36]

Figure 2.1: Literature search and selection process

## 2.2. Definitions for core concepts

### Automated decision making systems

**Automated decision making systems** (ADM) are socio-technical decision making systems anchored on the use of algorithm(s) that use quantitative analysis to ‘learn’ and apply the computational rationality to solve a certain decision task posed to the system (Spielkamp, 2019). ADM systems are based on computational logic however they do not have to be exhaustive in their capabilities, rather the term defines itself through the importance the respective decision algorithm has to the process of decision making. As such ADM systems can also entail human interaction and decision making.

A closely related concept to ADM are ’decision support systems’ (DSS). DSS gather and synthesize data to provide comprehensive reports to human-decision makers. The boundary between these to terms is fluid. Ultimately, we adopt the view that a system is an ADM system if the primary decision function is executed or dependant on the use of the algorithmic decision model. In such systems the direct or indirect influence of the algorithm is extensive enough to substantially influence the decision outcome.

### Ripple effects

**Ripple effects** as defined by Pel are effects that ”often start as a small problem in one specific place but cause more significant and often unforeseen problems further down in the system” (Pel, 2022). Contextualized to the social welfare system they are: ”the emergent amplified effects of social policy” (Pel, 2022). As such, ripple effects are a causal chain of self organizing behaviour between different actors (organisational, individual and infrastructural), triggered at a specific location with in a system of systems, resulting in compounded problems at other locations of that system. Ripple effects provide a description for the process of emergence of digital cages in social welfare, as multiple ripple effects can sum up to form a digital cage. They are therefore a contextualized description of the concept emergence.

### System Safety terms

According to Leveson a **System-state** refers to the specific condition or configuration of a complex system at a particular point in time. It includes the set of all relevant parameters, components, constraints and environmental factors that describe the system’s current situation. Understanding and analyzing the system state is crucial for assessing and managing safety because it allows for the identification of potential hazards and risks within the system (N. Leveson, 2011). **Constraints** are limitations or restrictions placed on a system or its components to ensure safe operation. These constraints can take different forms and are implemented to reduce risk and enhance safety. Constraints are an essential aspect of system safety and are used to manage and control various aspects of a system to prevent it from entering hazardous states or conditions (N. Leveson, 2011). A **Hazard** is defined as a potential source of harm or a situation that has the capability to cause damage, or adverse effects to people, its environment, or the overall system (N. Leveson, 2011). If a system is in a hazardous system state, this implies it is in an **unsafe system state**. **Harms** are the consequence of accidents or incidents occurring in a system that is in a hazardous state (N. Leveson, 2011). Harms can manifest in various forms and impact different aspects, including people, as well as the system’s functionality and integrity. The emerge through the combination of hazards with environmental worst case conditions. **Environmental conditions** surround a respective system, yet lay outside of the systems boundary of influence (N. Leveson, 2018). Meaning they can not be changed by the system but have to be accounted for. Ultimately, understanding and mitigating harms is a fundamental goal of system safety (N. Leveson, 2011). As previously explained, **control actions** or **-interventions** help to prevent harms by imposing safety **constraints** on system behaviour to ensure it operates within a safe state (N. Leveson, 2011). In this way system safety makes use of **hierarchy** theory. **Hierarchy**, facilitates the ”laws of behavior” at each system level. Those laws of behavior yield activity meaningful at a higher level. Hierarchies are established trough control processes operating at the interfaces between levels (N. Leveson, 2011). Complex systems are often consist out of an extensive hierarchy, which makes imposing control actions a difficult task. Control actions can be passive, meaning they maintain safety by their presence, or active, meaning some action to provide protection is necessary (N. Leveson, 2011).

## 2.3. Review Results

Appendix A provides an overview and summary of the analyzed literature.

From the analyzed literature it can be gathered that causes for the emergence of digital cages are numerous. Consequently the suggested solutions diverge as well. Therefore this literature analysis will subsequently list the in the literature suggested, most important, causes precipitating digital cages and go on to discuss the most relevant solutions as well as their underlying objectives suggested by the researchers. Prior to this short definitions of important concepts discussed in the literature are provided.

### 2.3.1. Relevant Concepts Discussed by Literature

**Administrative Burdens & Exclusion**

**Administrative burdens** refer to the time, effort, paperwork, and other non-financial costs imposed on individuals or organizations when they interact with government agencies or comply with government regulations. Administrative burdens can be associated with tasks such as applying for permits, licenses, or government benefits, paying taxes, or meeting regulatory requirements. Reducing administrative burdens in principal is an important goal for functional public policy and administration. **Administrative exclusion** is the exclusion of an individual or organization by a government agency from e.g. a service, market or (social) group. It can be the results of a specifically intended policy to regulate or govern, but it can also result out of excessive bureaucracy that overwhelms individuals or organisation subject to it. In the context of the digital cages, exclusion therefor can be a results of citizens unable to overcome administrative burdens placed on them (Peeters & Widlak, 2023). Consequently, there is ”good” exclusion that is necessary and wanted as well as ”bad” exclusion that produces harms and needs to be avoided.

**Red Tape**

According to Pandey and Scott Administrative **Red Tape** describes institutions (rules and procedures), that impose constraints on decision making (Pandey & Scott, 2002), they hence produce compliance burdens (Bozeman, 2000). As pointed out by Bozeman, in doing so they ”do not advance the legitimate purposes the rules were intended to serve” (Bozeman, 2000). They add unnecessary complexity that makes it more difficult to obtain information and make a ruling, thereby effecting the organisational performance of public administration, but also, as Peeters and Widlak point out, create burdens directly effecting citizens interacting with administrative bodies through, for example, creating catch 22 situations (Peeters & Widlak, 2018).

**Discretion**

**Discretion** refers to the degree of freedom that administrators have when interpreting, implementing and executing laws, policies, rules, and regulation. Barth and Arnold argue it is a core component of effective governance, as it represents the ”ability to exercise judgment as the situation changes” (Barth & Arnold, 1999). Discretion, if exercise properly, therefore can help administrators to customize decision making to the respective circumstances of each citizen. This can help to alleviate the burden placed upon those citizens.

**From street-, to system-, to infrastructure-level bureaucracy**

**Street-level bureaucracy** refers to the government officials directly interacting with citizens, such as case workers. The concept circumscribes the impact those street-level workers have on policy execution facilitated through their discretionary powers (Bovens & Zouridis, 2002). As Bovens and Zouridis explain **System-level bureaucracy** refers to a shift in the process of administrative decision making. Case workers are replaced or partially substituted by expert information systems. These systems supply the information, analysis and, in an increasing number of cases, the case decision. Thereby shifting discretionary power from the street-level bureaucrats to system designers (Bovens & Zouridis, 2002). Finally, **infrastructure-level bureaucracy** refers to the integration of such systems over various branches of government (Peeters & Widlak, 2023). In such a system of systems, the organisational boundaries already made fluid through system-level bureaucracy (Bovens & Zouridis, 2002), are resolved into one interconnected network structure (Peeters & Widlak, 2023). Street-level discretion is further reduced (Peeters & Widlak, 2023) and transfused into ’digital discretion’ (Bullock, 2019). The overall system boundaries have become increasingly vague, resulting in a more diverse group of interacting subsystems and actors. The system exhibits complex behaviour.

### 2.3.2. Precipitating Causes

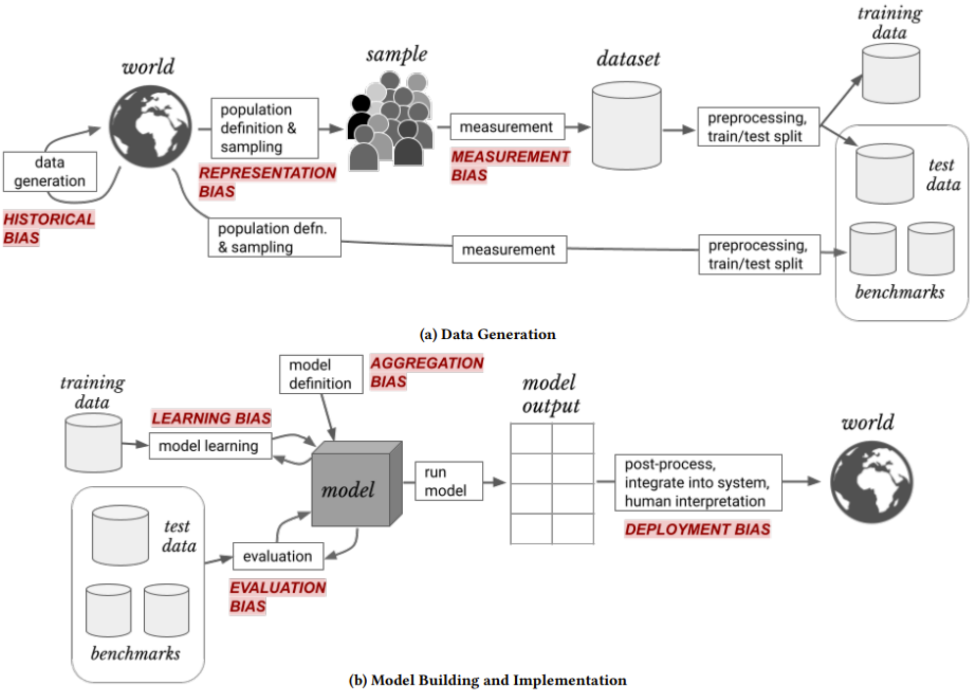
Researchers Peeters and Widlak initially introduced digital cages as ”a highly disciplining infrastructure that rationalizes the execution of tasks through information architecture and algorithms instead of Weberian rules and procedures” (Peeters & Widlak, 2018). In consequence Peeters and Widlak characterize a digital cages as a system: (1) that perpetuates mechanisms of unreasonable exclusion, (2) that presents a black box to citizens and bureaucrats leading to a lack of oversight and control over decision making, (3) that produces ‘legal contamination via ICT’ by forcing its characteristics onto operational execution, (4) that eliminates street level discretion and (5) that creates perverse behavioural incentives for administrative bodies to take action against their citizens (Peeters & Widlak, 2018).

These characteristics are a consequences of the structural logic underlying public administration (Peeters & Widlak, 2018). Researchers have identified several factors that contribute to this dynamic. Firstly, many researchers point towards the structural rigidity of information architecture in public administration as a cause for the emergence of digital cages (Jorna & Wagenaar, 2007) see also (Veale & Brass, 2019). They argue that the existent information architecture often inhibits organisational learning and hence obstructs operational or ’street-level’ discretion (Jorna & Wagenaar, 2007). Furthermore, researchers argue that AI aided ADM systems are also subject to set of ridged rules and hence will lack discretion in situational dynamic ’boundary cases’ (Barth & Arnold, 1999). On the contrary however they cite a future powerful AI may reach the ability to exercise judgment under fluid situational circumstances (Barth & Arnold, 1999). Nonetheless as Bullock points out: ”even if the overall quality of administration can be improved, AI changes the nature of risks to good governance in significant and important ways” (Bullock, 2019). Consequently, an increase in AI capability does not necessitate an improvement of human-centered discretion in public decision making. Additionally, strong AIs may still be anchored in a rigid IT infrastructure that hinders its ability to exercise discretion. This is underlined by Lorenz et al. who admit that the increased standardization of work caused by ADM systems and the IT infrastructure its deployed in obstructs the utilization of ”local knowledge”. It can be argued that this ”centralized rationality” ultimately results in irrational deployment of ADM (Lorenz et al., 2021). This argument is in line with fundamental critiques on centralized forms of decision-making (Bungay, 2021). Importantly however, ”digital” and ”discretion” are however not dichotomous (Ranerup & Svensson, 2023) but, rather, depend on interplay of ”technologies, their design and emergent use” (Ranerup & Svensson, 2023). Consequently, the ”digital characteristics” in itself are not perpetuating digital cages, rather it is the manner in which they are utilized within the current system.

The second factor numerous papers have explored is coined as ’flawed data categorization and labeling methods’ utilized in public information architecture (Peeters & Widlak, 2023). Examples include the collection of information that misrepresents citizens leading to consequential generalizations such as presented by Widlak and Peeters (Widlak & Peeters, 2020). Yet flawed labeling and categorization also entails using data without sufficiently eliminating historical biases present in the data sets (Weyerer & Langer, 2019). Additionally it also refers to issues of ’optimization’ and ’generalization’ often encountered in machine learning. Mulligan and Bamberger point out that algorithms are trained to ”optimize over large sets of data”, consequently they are susceptible towards missing ”distinct patterns in small subpopulations” (Mulligan & Bamberger, 2019). Such misclassification inadvertently leads marginalization of these communities. Closely connected to the aforementioned issues are the insufficient data quality and completeness often encountered in public administration systems. According to Pel this misrepresentation can often lead to ’ripple effects’ effecting citizens in ways that in hindsight are no longer comprehensible (Pel, 2022). Another often cited reasons for the emergence of digital cages are bureaucratic rules and procedures that perpetuate slow communication between and within administrative organisations (Jorna & Wagenaar, 2007). Those rules ultimately hinder organisations to dynamically as well as effectively tackle arising problems in development and operations.

Among the perpetuating factors for digital cages are also privacy and data protection concerns (Peeters & Widlak, 2018). Often these concerns, which follow valid reasoning, hinder governmental organisations to share information between each other, ultimately enabling the emergence of the aforementioned ’ripple effects’ (Pel, 2022).

Finally researchers often point to systemic biases and discrimination that are present in public ADM design (Wirtz & Müller, 2019) see also (Weyerer & Langer, 2019). Such biases can have various reasons ranging form data pre-processing biases such as historical biases in data sets to learning biases in ADM training. As well as the often overlooked issues of deployment biases, referring to miss use of an ADM system in deployment. Suresh and Guttag have provided a exemplary categorisation of the different biases through the design process of a machine learning model (Suresh & Guttag, 2021) that can function as a guidance in understanding the challenges encountered during ADM systems design. Figure 2.2 outlines these different sources of biases. In AMD system for governance literature, there seems to be a strong focus on mitigating 'pre-processing' and 'in-process' biases, illustrated in Figure 2.2 as historical-, representation-, measurement-bias and learning-, evaluation-, aggregation-bias, respectively. The also pointed out component of 'post-processing' exemplified by 'deployment-bias' is often overlook. Additionally, the literature seems to overlook performance issue of ML or AI driven ADM systems during operations. This fact will be analyzed in the following chapter.



(a) Data Generation

(b) Model Building and Implementation

Figure 2.2: Biases in Machine Learning Applications Development (Suresh & Guttag, 2021)

Biases often also result as a functions of the developer teams demographics and individual characteristics (Weyerer & Langer, 2019). This aspect is closely related to the power dynamics surrounding the design of ADM systems in public administration. The power structures dominating ADM projects have a significant impact on the final design outcome (Alkhatib, 2021). As Graycar and Masters illustrate such power dynamics, if not kept in check, can lead to very severe cases of harm (Graycar & Masters, 2022).

Aside form structural components and general design process challenges regarding ML or AI driven ADM systems researchers have also identified public design specific issues, that lead to the emergence of digital cages. Nouws et al. show that inadequate "public design practices" are a primary reasons for the emergence of digital cages (Nouws et al., 2022). They illustrate that public design processes of algorithmic systems: "(1) are often narrowly focused on technical artefacts, (2) disregard the normative basis for these systems, (3) depend on involved actors' awareness of socio-technical components and interactions in public algorithmic systems, (4) and are approached as linear rather than iterative" (Nouws et al., 2022). The growing literature surrounding the topic in question underlines the discrepancies that can be observed between the ADM systems in design and in their deployed state. Veale et al. point out that many of these discrepancies surface in later stages of the respective innovation projects, when a design choices is firmly restricting the outcome (Veale et al., 2018). They go on to explain that currently implemented design methods fall short of providing sufficient transparency and adaptation space (Veale et al., 2018). Others point towards a distilled lack of a governance framework for ADM design (Jonk

& Iren, 2021). Affords to derive such a framework have been made (Wirtz et al., 2021). Additionally, researchers have put forth complementary frameworks to assist in ADM design. Such as frameworks for risk management (Bannister & Connolly, 2020), technology impact assessment (Ojo et al., 2019), as well as frameworks to determine automatizing potential (Etscheid, 2019) see also (Young et al., 2019). However it remains unclear if these frameworks are currently in use and what their respective utility is. Moreover, none of the frameworks is derived using a holistic system safety perspective, they rather focus on specific design objectives, such as ”participatory” or ”accountable”, however do not explicitly mention system safety as an objective.

### 2.3.3. Objectives & Solutions

The above presented findings suggest that the causes for the emergence of digital cages are of systemic nature and hence any attempt of solving them can not be one dimensional. In accordance with this fact researchers have defined a number of objectives that could help preempt digital cages, and based on those objectives discuss possible solutions focused on preventing or mitigating digital cages. The arguments are summarized as follows.

Loi and Spielkamp highlight the importance of accountability in the use of AI by public administrations. The authors emphasize the pivotal role of transparency in algorithmic systems, as well as the need for mechanisms for auditing and challenging automated decisions(Loi & Spielkamp, 2021). There arguments are in accordance with the principals for AI design Floridi and Cowls coined as ‘explicability’ (Floridi & Cowls, 2019). According to Karkliniewska administrative transparency guidelines need to entail the three main pillars of the public sector functioning, specifically: (1) ”the implementation of policy and law enforcement”, (2) ”the organizational environment” and (3) ”all mechanism of the ongoing process of used models” (Karkliniewska, 2022). Le Dantec and Edwards add that, in practice, mechanisms of accountability are notoriously hard to implement due to the fact that public administration information systems cross over different ”scales and chains of accountability” (Le Dantec & Edwards, 2010) see also (Veale et al., 2018). Veale et al. conclude that ADM systems therefore need to be ’internally accountable’ (Veale et al., 2018). On the contrary however concerns that transparency as a function of accountability might not always increase public trust and buy in are also being raised (Veale & Brass, 2019). Specifically, cases or projects centered around ‘taboo trade-offs’ may diminish the public legitimacy when transparency is applied (Veale & Brass, 2019). Furthermore, issues with providing public explanations may infringe on the ’transparency’. Providing the public with in depth explanations on how ADM systems work may lead to the ”gaming” of such algorithms by harmful actors (Edwards & Veale, 2017). Additionally, local explanations are also likely to cause confusion and hence fall short of providing ’meaningful explanation’ to users (Edwards & Veale, 2017). Finally Janssen and Kuk point out that ”algorithms dynamically co-evolves with data, systems and humans within complex socio-technical system” (Janssen & Kuk, 2016) this ever evolving dynamic can not be encapsulated within a local explanation, hence possibly further reducing its meaningfulness. Nonetheless, XAI can help to improve transparency by offering an interactive means to citizens. Good explanation of automated governmental decision have shown to increase the citizen’s willingness to accept an administrative decision (van Engers & de Vries, 2019). Scholars agree that ’accountability’, ’transparency’ and ‘explicability’ are fundamental design objectives for public ADM systems and are necessary in order to prevent the emergence of digital cages. In any case this requires the system and is state to be ”measurable”, since only information that is known can act as a baseline to insuring these design objectives.

Filgueiras underscores the significance of human-centered design in AI systems. By involving diverse stakeholders in the design process, including those affected by algorithmic decisions, it is possible to address biases, ensure fairness, and reduce the likelihood of digital cages (Filgueiras, 2022). The aforementioned argument represents one of the core concepts put forth in the literature. Several authors endorse a broader, interdisciplinary and inclusive design approach in order to mitigate the emergence of digital cages, see (Veale et al., 2018), (Wirtz et al., 2020) & (Young et al., 2021). Additionally, a multi-actor design approach can also aid to ”curb the curse of flexibility” (R. I. J. Dobbe, 2022). The ’curse of flexibility’ first introduced by Leveson applies to systems that are no longer limited by physical or materialistic constraints but rather by human intellectual capacity (N. Leveson, 2011). Collington underlines the importance of this issue for the welfare state she views the past privatisation of public sector digitization as an infringement on public normative guidelines for ADM system development. And therefore urges ”to reconsider the governance and ownership relationships in public sector digitisation to ensure they are able to steer technological development in ways that benefit citizens and society” (Collington, 2022).

Djeffal proposes normative guidelines for AI in government and public administration. By integrating ethical principles into the development of algorithmic systems, policymakers and practitioners can mitigate the risks of digital cages and ensure that AI is used to serve the public interest (Djeffal, 2020). A lack of ’normative basis’ has also been criticized by Nouws et al. They argue that unclear normative guidelines obstruct the designers ability to anticipating and address possible unintended consequences within the design process. Consequently, Nouws et al. call for a more structured deliberation on ”univocal and holistic normative or evaluative frameworks for public algorithmic systems” by politicians and decision-makers that is not guided by reactive practices (Nouws et al., 2022). Such a deliberation however would require accurate feedback mechanisms instilled within the design process.

Van Noordt and Misuraca emphasize the need for robust impact assessment frameworks for AI technologies in public services. By evaluating the social, economic, and ethical implications of algorithmic systems, policymakers can proactively address the potential harms and unintended consequences that contribute to the creation of digital cages (van Noordt & Misuraca, 2020). Authors have suggested that positive effects of the technology such as increasing scalability, decreasing cost, and improving quality must be balanced with the potential concerns surrounding artificial discretion, equity, manageability, and political feasibility (Young et al., 2021).

The majority of the collected literature analysis the problem from a deterministic perspective, by focusing on a narrow subsection of the system. However as Nouws et al. point out, this approach exacerbates the focus on technical problems and fixes surrounding the phenomena (Nouws et al., 2022). Dobbe points out that such a perspective for ecosystems involving AI and ADM systems is especially problematic, as the search for ”causality” results in a narrow focus on technical factors, engineering activities and operator errors, thereby overlooking systemic factors(R. I. J. Dobbe, 2022). He however suggests that ill defined system boundaries and ambiguous system states are often at the root of any accident (R. I. J. Dobbe, 2022).

In the gathered literature the digital cage phenomenon is discussed form various perspectives. Researchers predominantly discuss issues viewed as root causes for the emergence of digital cages as well as abstract public objectives that will likely help to alleviate the problem. Such advise or suggestions as pointed out by Veale et al. are often abstract in nature or are accompanied by assumptions that may not always be applicable to practice (Veale et al., 2018). Nonetheless, many point towards flawed design practices within public administration as a dominant reasons for the emergence of digital cages. Correspondingly, many of the mentioned public objectives or values such as transparency, accountability, participation, visibility and auditability etc., seek to increase control over the design process to make errors more explicit. Many hint towards a desire to move to an iterative design process. The design process however is rarely dissected within the literature. This circumstance represents a gap in research, since it is unclear as to how the emergence can be prevented or mitigated at different stages of the design process and how the suggested public objectives should be incorporated into set process.

## 2.4. The Role of Deployment & Operations: A System perspective

Though there are numerous reasons mentioned, a majority of scholars agree on the pivotal role of design practices in the precipitation of digital cages. Conversely several scholars argue for a shift towards an iterative design methodology. They argue that a shift to a participatory, iterative design methodology is necessary in order to enable better detection and correction of biases as well as general design flaws. Crucially however, there seems to be a heavy emphasis on the early stages of the design, while few address issues in the area of system-deployment, -operations. Many researcher focus on high level design objectives such as ”participatory”, ”human-centered”, ”accountability” and ”robustness” (Filgueiras, 2022), also (Loi & Spielkamp, 2021), also (Djeffal, 2020). Additionally they emphasis the importance of introducing frameworks and methodology to asses risks associated with ADM-systems (van Noordt & Misuraca, 2020), however few mention the need for a holistic systems safety approach. Almost all authors focus on the importance of design, while few analyse the role deployment and operations within any given design methodology. Given the emergent nature of digital cages, hence their non-obviousness, this focus may prove to be presumptuous. Digital cages can be characterized a emergent phenomenon in a complex system. Such systems exhibit chaotic characteristics, which make them extremely sensitive to system state conditions. Hence small system status discrepancies between the early stage design environment of an ADM system and the deployment/operation environment can trigger unforeseen skewed and distorted outcomes at aggregated system levels. Consequently, while sound design practices in the initial state of an ADM project are necessary they do not infer that the system once deployed will behave as intended.

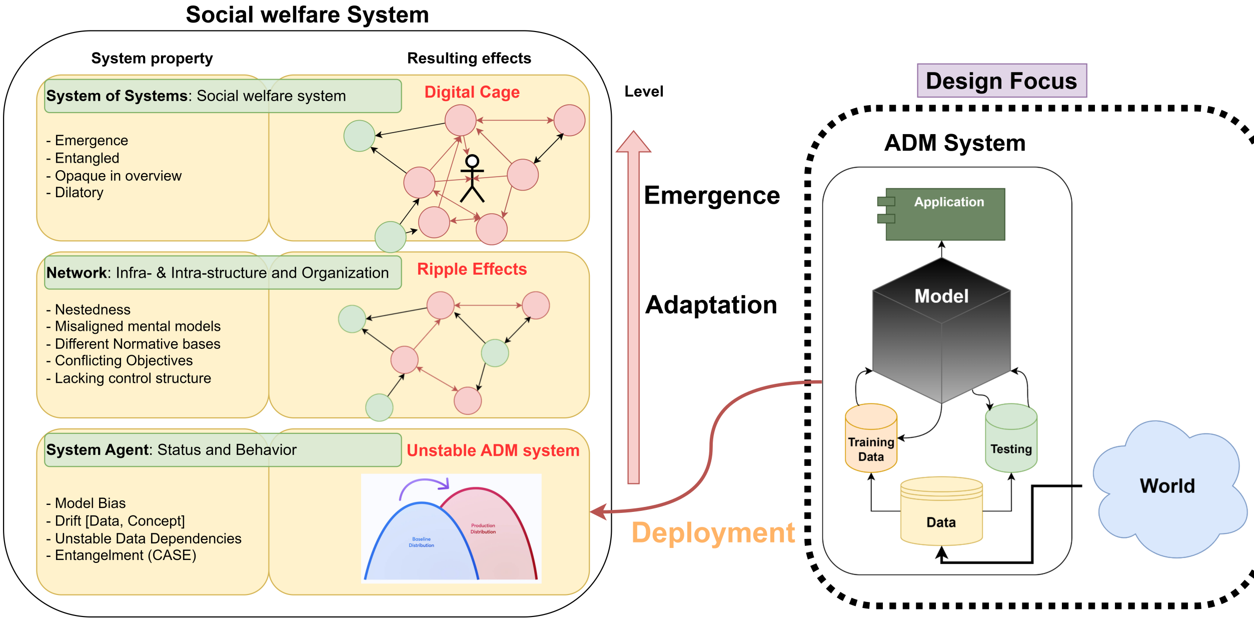


Figure 2.3: ADM in the Welfare System of Systems

Figure 2.3 exemplifies this relation. It illustrates the ADM system as a system within the larger social welfare system. Often the focus is directed towards the ADM system itself, however this unit of analysis is to reductionist. By limiting the analysis to the ”System Agent” we loss the interaction patterns with other system components, which makes it impossible to identify ripple effects and the there-out resulting Digital cages. A pure focus on ADM design methodology therefor will often fail to mitigate the digital cages, because within in that perspective the existing causal relation can not become apparent. Such a perspective neglects two major sources of possible hazards in the development process: system deployment and operations.

As outlined in Chapter 1 both ”Deployment” and ”Operation” represent ”Adaptation” which can trigger ’emergence’. In fact the represent the two processes through which the emergence of a digital cage may be triggered. In deployment and operation the system is integrated within its larger system of systems context. During prior design stages this is not the case. The next sub-chapters will outline, how in each of those cases a digital cage may emerge and relate these observations back to the previously outlined design objective from researcher.

### 2.4.1. Deployment: A safety Gatekeeper

Deployment represent process through which the ADM system is implemented into it larger system landscape. This process can evoke the emergence of a digital cage either through implementing an inherently flawed ADM system into the larger system of systems or through implementing an ADM system not fit for purpose. Inherently flawed ADM systems are systems that intrinsic mistakes, such as pre- or in-processing biases. While systems not fit for purpose can be categorized as ’deployment bias’. The latter emerge through discrepancies between the design and deployment environment.

Deployment therefore represents the last line of defence against the implementation of bias ADM systems into the larger system landscape. Current practices within this process however focus on the technical implementation and integration of the new ”System Agent” into its technical landscape. Developers are focused on technical performance indicators e.g. outages. They fail to capture and analyse the triggered changes in system behaviour, that resulted out of the implementation of a new system component. As a results, technically sophisticated ADM systems are deployed in unverified likely unsafe system states, with highly limited tractability of any possible future safety insurgence. Literature surrounding complex systems engineering has shown that sound deployment methods are key in order to adopt a truly effective system intervention (Brazier et al., 2018). Furthermore sound deployment practices represent a prerequisite to the successful adoption of an iterative design methodology (Myrbakken & Colomo-Palacios, 2017). Additionally, more and more practitioners and researchers agree that while deployment practices must consider general best practices of system deployment they must hold to equally high regards safety objectives (Myrbakken & Colomo-Palacios, 2017). The issue of ADM system deployment can be viewed as a control action problem. The implementation of faulty or unfit for purpose ADM systems can be traced back to a lack of detection, interpretation and recovery measures in place for the deployment process. Such control interventions could help to ensure ’accountability’ & ’tracability’. Ultimately, sound deployment practices also evolve around enabling ’feedback’, as any encountered problems or challenges must be revised and rectified in subsequent design cycles.

### 2.4.2. Operations: Managing System State

Operations differs from Deployment in that it is a continuous process. In Operations the system operates in a fluctuating environment that evolves around ’Adaptation’ of system state. Ultimately, allowing the system to migrate to a higher risk level (Rasmussen, 1997), as previously outlined in Chapter 1. In practices this is a process is triggered through a indented change of a neighboring subsystem to the ADM system or through a number of small changes in the operating structure over time. Often the system experiences deterioration through unsupervised divergence from its indented operating process through for example changes in data entry practices by operators, that results in the quality of data deteriorating over time. If a predictive component is involved leading to a quality decrease in decision making that is difficult to spot. Indented changes my represent the second source of harm. Such may be the implementation of an interdependence with a new subsystem that utilizes the manufactured output of the ADM system for purposes it was initially not intended for. In such cases overconfidence in the models predictive capabilities leads to wrongful decisions. Further examples are outlined in (Sculley et al., 2015), (Côté et al., 2023) (Breck et al., 2017). Like deployment, operations also represents a control action problem. That can be improved through clear detection, measurement, interpretation and recovery protocols. By improving the ability of operates to observe whether a system still operates in a ’safe state’ one is able to mitigate the emergence of a digital cage. Hence control and feedback are pivotal to ADM systems safety in social welfare.

### 2.4.3. Control, Communication and Constraints: the Backbone of Safety

For the particular application of safety control interventions focused on the prevention of digital cages, the importance of safety centered design can not be overstated. Dobbe et al. underline the need for a holistic system safety approach to ADM system design (R. I. J. Dobbe, 2022). As Leveson points out, safety is an emergent property meaning: ”Safety can be determined only in the context of the whole. Determining whether a plant is acceptably safe is not possible, for example, by examining a single valve in the plant.” (N. Leveson, 2011). As Leveson explains this in turn dictates that safety of an artifact or system can not be established without information about the ’context’ in which this system is utilized (N. Leveson, 2011). Revered back to ADM systems, the safety of such systems can only be determined through the relationship between the ADM system and other system components.

To ensure safety ’control’ over the system must be exhibited. To exhibit effective control several conditions must be meet, as pointed out by Leveson (N. Leveson, 2011), see also (Ashby, n.d.):

> ”In order to control a process, four conditions are required:
>
> - Goal Condition: The controller must have a goal or goals (for example, to maintain the setpoint).
> - Action Condition: The controller must be able to affect the state of the system. In engineering, control actions are implemented by actuators.

- Model Condition: The controller must be (or contain) a model of the system.
- Observability Condition: The controller must be able to ascertain the state of the system. In engineering terminology, observation of the state of the system is provided by sensors.”

Within a system of systems another component pivotal to execute control is ’communication’. Communication enables the proper allocation of information and knowledge, as well as feedback for system improvement.

Fundamental to the the safety of a system is however the ”constraints” higher level system components place onto the system (N. Leveson, 2011). **Constraints** ensure the system is not operated in unsafe conditions through restricting the emergent properties of a system. This prevents the occurrence of unsafe behaviour or is mitigates its potential for harm. **Constraints** are passed down unto a system form a higher level of control. Meaning an actor also called controller that has the authority to invoke control. Figure 2.4 taken from the STPA Handbook (N. Leveson, 2018) illustrates this dynamic.

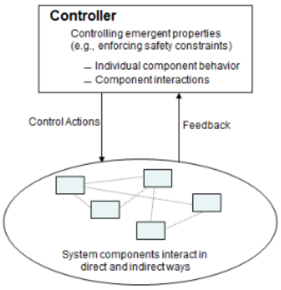

Controller
Controlling emergent properties
(e.g., enforcing safety constraints)
– Individual component behavior
– Component interactions
Control Actions
Feedback
System components interact in
direct and indirect ways

Figure 2.4: Constrains enforced through controller onto system (N. Leveson, 2018)

Constraining the systems behaviour is vital to its safe operation. Due to the increased risk of emergence, this holds especially true for complex systems. To invoke these constraints effectively, one needed the previously mentioned attributes of control and communication. Additionally, one need an effective and concrete method of invoking these constraints. This can be characterized as a safety control intervention.

The aim of this research is to identify those safety control interventions using a holistic system safety approach. The control interventions should entail detection, measurement, interpretation/diagnosis as well as response/recovery instruments facilitated through effective communication. Thereby enhancing the overall system safety and mitigating the risk of digital cage emergence.

## 2.5. Conclusion & Research Question

The literature analysis has also reaffirmed that social welfare is one area increasingly under threat of the emergence of extremely harmful forms of digital cages. This development is driven both by the proven interest of governments to increase efficiency and effectiveness in this area, resulting in a higher number of focused digital policy efforts compared to other policy areas. And secondly due to the general nature of administrative decisions concerning social welfare, arising harms are particularly harmful to individuals. Often the individuals are also part of a very vulnerable or marginalized groups of society. Research into the prevention of digital cages in this area is therefore most pressing as Alston eluded to (Alston, 2019). Therefore, this research will focus on the subject of digital cages in social welfare, with a particular focus on the Netherlands, due to the authors affiliation with the Netherlands. The literature review has also confirmed the authors suspicion regarding a clear lack of research for safety control actions and interventions during system deployment and operation, who take a system of systems perspective. The review shows that researchers are focused on high level top-down design objectives as well as on the specifics of ML and AI driven ADM system design. Yet many fail to recognize that structure can form its own values. Hence they fail to see that by implementing a new ADM system into larger system of systems the structure is modified and new value and properties arise. Based on this however a constraint based approach of ensuring safe during deployment and operations can help to prevent the emergence of unwanted behaviour in that specific system of systems. Designing safety control interventions that take this into account, can therefore yield good returns for both. Researchers understanding of ADM systems in a larger system landscapes. As well as practitioners ability to mitigate harms through an improved understanding of system safety practices.

To maximize the impact of this research, the focus will be directed towards ADM initiatives on the national level. The study of the phenomena on the national rather then a regional or municipality level is likely to improve the overall system safety of social welfare to a larger degree. Additionally, studies at this level may be more generalizable towards other countries or contexts. During the study of the literature it became apparent that for the purpose of this thesis an analysis of the previously described ”toeslagenaffair” represents a good point of departure. The ”Toeslagenaffair” represents a good unit of analysis, because it shows a clear emergence of a ’digital cage’, has accumulated a large body of documentation and illustrates interesting dynamics between different actors that changed over time.

Consequently the following research question was derived to meet the identified research gap:

**”What are safety control interventions to curve the emergence of algorithmic decision making systems-induced digital cages in the context of national social welfare administration in the Netherlands?”**

# 3 Methodology

The following chapter aims to derive an appropriate methodology for research of the previously proposed main research question. It initiates the argumentation with the main research approach before moving on to cover the research sub-questions and their respective methods. Each sub-question will be introduced with a discussion about the expected output the question may yield as well as relevant knowledge and data needed to answer the question.

## 3.1. Research approach

Building upon the central research question outlined earlier, the primary objective of this research is to demonstrate the utility of safety control interventions based on system safety theory for ADM systems deployment and operations within the Dutch social welfare system. The research aims to provide practitioners with tools to preempt the formation of digital cages, through enabling them to enhance their systems safety for the deployment and operation of ADM tools. The control actions and interventions should be comprised out of detection, prevention as well as mitigation instruments that can be utilized during the deployment and operations process against digital cages. To these ends the research will follow a design science approach. Design science is particularly suited to create and evaluate artifacts (Hevner et al., 2004). Artifacts can be construct, model, method, and instantiation (Hevner et al., 2004) therefore a set of control interventions if represented comprehensively satisfies this definition. Design science research is commonly separated into different steps Peffers et al. define them as: (1) Identify Problem & Motivation, (2) Define Objectives of a Solution, (3) Design & Development, (4) Demonstration, (5) Evaluation and (6) finally Communication (Peffers et al., 2007). Step 2 to 5 can be processed in an iterative fashion. Each step posses an important intrinsic question to this research project, hence has to be addressed within a sub research question. The sub research question therefore tie into the chosen research approach. Step 1 is a prerequisite to this research project, it is addressed though the introduction. However in order to sufficiently address subsequent design steps the problem must be analyzed in depth. Only after having sufficiently segmented the system in question can reasonable objective be derived. Therefore the first sub research question will address the in depth analysis of the system state of an exemplary ADM system. Design step 2 is closely related to step 1 as the definition of the problem ultimately dictates objectives for the solution. It will be addressed with the second sub research question. The objectives however are subject to change and may be altered during subsequent design iterations. Step 3 the 'Design & Development' is central to the research and will be conducted using the framework provided by Hevner et al. The framework uses an iterative 'assess' and 'refine' process that takes both the "environment" and as well as the "knowledge base" into account, also known as the "Relevance" and "Rigor" cycles (Hevner et al., 2004). To this purpose step 4 is integrated within to this process, since without a demonstration assessment and refinement of the design are not feasible. By conforming to this methodology it is ensured that the resulting design artefact is both grounded within existing theory, yet still applicable in the context of the real world environment. Applied to the specifics of this research subject this implies that the design addresses the social welfare environment present in Netherlands as well as incorporates existing system safety theory such as Leveson and Dobbe (N. Leveson, 2011) also (R. I. J. Dobbe, 2022). Step 5 can be addressed within one research question provided previous ex ante evaluations are considered in this chapter as well (Hevner et al., 2004). Finally ”communication of the design” will be addressed during the reflection of the research project itself. Since communication represents an external objective of this research it is not ascribed with a sub research question. Nonetheless, the final section of the research deliverable will assess the degree of generalizability of the research findings and propose a suitable communication strategy.

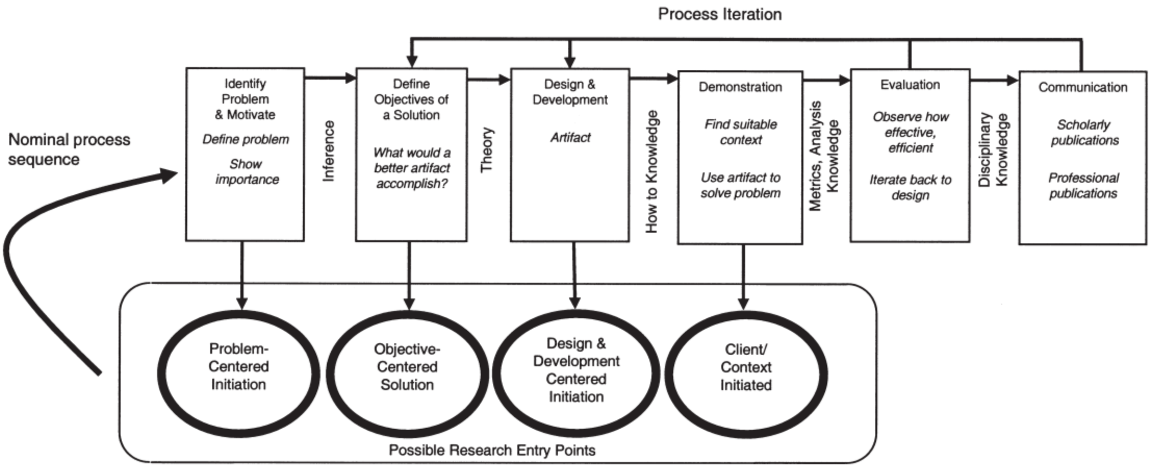

Process Iteration
Nominal process sequence
Identify Problem & Motivate
Define problem
Show importance
Inference
Define Objectives of a Solution
What would a better artifact accomplish?
Theory
Design & Development
Artifact
How to Knowledge
Demonstration
Find suitable context
Use artifact to solve problem
Metrics, Analysis Knowledge
Evaluation
Observe how effective, efficient
Iterate back to design
Disciplinary Knowledge
Communication
Scholarly publications
Professional publications
Problem-Centered Initiation
Objective-Centered Solution
Design & Development Centered Initiation
Client/ Context Initiated
Possible Research Entry Points

Figure 3.1: DSRM Process Model, Proposed by Peffers et al.2007

## 3.2. Research Questions & Deliverables

### 3.2.1. SQ.1: System State & Problems

Identifying the right requirements and objectives for the design of the safety control interventions is pivotal to the results of this research. However identifying meaningful objectives and requirements is only possible if the problem, its constraints and resulting hazards, are identified and analyzed. As such this research is in need of a exemplary case that can be used in order to conduct a description of the current system state and an analysis of constraints and hazards within the system. While the literature review above provides a first motivation of this research it is not extensive and in depth enough in order deduct requirements and objectives for safety control interventions. This analysis must therefore be addressed within the research on an exemplary case, with the following research question:

**SQ1: ”What hazards in the socio-technical environment enabled the emergence of a digital cage during the ADM systems use in the dutch welfare administration at the time of the Toeslagenaffair?”**

The question recognizes that the problem must be analyzed in its socio-technical context and seeks to answer important questions needed in order to derive subsequent design objectives and deliverables. For example: Which stakeholders are involved and what are there power relations?, What policy aspects influence the environment?, What technology & process standards exist and are they applied? What is the existing safety control structure and is it safe? To this purpose the ”Toeslagenaffair” represents a good unit of analysis for this research, as previously outlined in chapter 2. In order to derive this necessary insight, data about stakeholders, systems state, utilized processes and technology, as well as theoretical knowledge must be gathered. The question will be answered by combining two aspects. Firstly, the analysis will focus on the different stakeholders involved and map respective needs and wants, as well as power. To this purpose the research will utilized a structured document review process focused on identifying relevant information in the fast amounts of documents related to the Toeslagenaffair. The process implemented for this purpose is outlined in the methodology section 3.3.2. This part of the analysis will enable us to identify the environmental conditions present at the time of the toeslagenaffair. Secondly, the analysis will utilize aspects of system safety theory, STPA (Systems- Theoretic Process Analysis), to derive the safety control structure existent in the Belastingdienst at the time. This is done to identify potential hazards and constrains as well as map the institutional environment. The gathered knowledge will be combined into a comprehensive overview of the current system state structured into technical, social and institutional domain. Finally, this information will be gathered in the System Theoretic Process Analysis (STPA) of the system. With the help of STPA the hazards within the system will be identified. These hazards will subsequently be used in sub question two, to derive system safety requirements for the safety control interventions.

### 3.2.2. SQ.2: Solutions Objectives & Requirements

The second sub research question is intended to facilitate an analysis that allows to derive a set of objectives and requirements for the subsequent design of the control interventions. As stated prior the objectives must be inferred rationally through the problem analysis (Peffers et al., 2007). Therefore the sub research question is formulated at follows:

**SQ2: ”What are objectives and requirements for ADM system safety control interventions to curb digital cages in dutch social welfare administration, derived from the example of the toeslagenaffair?”**

The objectives and requirements are subject to iterative refinement and therefore may change during the design process. Despite this the requirements should be set for the design artifact as specific and early as possible, since they ultimately inform the design process (Peffers et al., 2007). The the derived objectives and requirements should be derived by finalizing the STPA analysis previously initiated through sub-question one. This does not infer that all requirements will be quantitative. Due to the nature of the design artefact it is likely that a majority of the objectives will be qualitative. Constraints however should be stated as explicitly as possible. The final deliverable of this research question is a full categorization of an objectives and requirements list. Finally, some general requirements regarding the intervention formulation and structuration will be derived based on system safety theory.

### 3.2.3. SQ.3: Design & Development

The design and development of the safety control interventions represents the core aim of this research. Determining the structure and functionally of these control interventions is dependent on the objectives and requirements in sub question two. Ensuring the applicably of the control interventions is an intrinsic objective of this research that yields great insight into their utility. Therefore the third sub research question can be framed as follows:

**SQ3: ”How can safety control interventions for ADM systems in dutch social welfare be structured, formalized and applied in order to adhere to the identified objectives and requirements?”**.

Subject of this sub question is the design of the safety control interventions. Particular care will be directed towards the structure and formalization of the control interventions to allow for a high degree of utilization and applicability of the interventions to future ADM system related projects in social welfare. Additionally is it the aim of the safety control interventions to yield insight in to potential system structure adjustments that are likely to improve the prevention and detection capabilities for digital cages on the example of the toeslagenaffair. Hence, part of this sub research question entails that the control interventions be underlined with a concrete demonstration (Johannesson & Perjons, 2014) of an intervention or system control mechanism that can help to enforce the safety control interventions. The system landscape those interventions should be applied in is complex. It therefore is subject to constant change and adaption. The resulting intervention structures therefore must be formalized with a particular focus on robustness. Consequently, the deliberation of the design will be conducted with the knowledge gathered from literature. System safety theory builds the ground support for the safety control interventions. By grounding the control interventions in this theories their safety relevance is insured and valuable evaluation time is retained. An additional source for safety control interventions will be work conducted in the field of ”DevOps” and ”DevSecOps” (Myrbakken & Colomo-Palacios, 2017). These concept represent the latest developments in iterative design methodology, which are utilized be industry leading technology companies. Though those methods are likely not applicable in their current form the underlying principals governing these methodologies will likely yield insightful information for the intervention design.

### 3.2.4. SQ.4: Demonstration & Evaluation

Design science affords the crustal opportunity to evaluate the derived principals with empirical and qualitative methods(Hevner et al., 2004). This is of particular importance since evaluating the principals necessitates an analysis within the given organisational context (Hevner et al., 2004). Evaluating the principals with a mix approach consisting out of both empirical as well as qualitative methods enables to insure a high level of relevance as well as rigor. Such an approach however can be extensive choosing the right evaluation methods is therefore critical to this research. Appropriate methods must both, greatly inform the design itself, as well as clearly outline the utility presented by the design artefact. Due to the nature of the design this research is only able to analyse the interventions ex ante (Johannesson & Perjons, 2014). Consequently, the design will be evaluated with the help of a workshop series. As such the final sub question is formalized as follows:

**SQ4: ”What is the utility of the safe deployment principals for ADM systems in dutch social welfare?”**.

Answering the utility of the safety principals is the main objective of this research question. Data provided through the workshop series will be of direct implication for the applicably of the designed interventions. During the workshop series practitioners with different background will discuss and evaluate the designed interventions. Conversely this analysis will yield insight into the effectiveness of the interventions in preventing hazardous system circumstances that may precipitate digital cages. Nonetheless, the utility can only be addressed moderately within this research project. Ultimately, only the practical utilization of the designed control interventions during an ADM system deployment and operations would provide an indefinite answer to this question. Such an ex post evaluation however is costly in precious resources and on the other hand due to the importance of getting ADM deployment and operations ’right’ the interventions can not be employed in such a capacity without prior exploration that hinds at their potential utility for the system in question. Consequently, the chosen ex ante evaluation technique provide a good compromise that recognizes both this instance as well as the time limitations imposed upon this research project.

## 3.3. Methods

This section will outline the main research methods alluded to in the different sub research questions. A particular focused on justifying the choice for STPA analysis for the system safe analysis.

### 3.3.1. Evaluation Workshop

This thesis will make use of a workshop series to evaluate the design artefact/interventions. The purposed process of the workshop can be seen in Figure 3.2. Several reasons for use of a workshop series as validation method, can be put forth, as pointed out by Thoring et al. Firstly, workshops allow for conducting evaluations in a comparable and replicable way, which helps improve research rigor in the design science field. Secoundly, they are effective in co-creation & evaluation of artifacts in naturalistic settings. This is particularly useful in formative evaluations where the goal is to test and improve the artifact in the context of its use. Finally the put forth that workshops facilitate direct engagement with stakeholders, including users, designers, and other relevant parties. This engagement is vital for gathering feedback, understanding stakeholder needs, and ensuring that the artifact’s development aligns with those needs (Thoring et al., 2020). For this thesis the workshop will be conducted through a moderated online ’Miro’ session. The workshop will approximately take 1 hour with around 5 participants per session. The sessions will be split in manner that reflects the different backgrounds of the participants. This enables participants to speak a ”shared” language allowing for deeper discussions of the interventions form their expert perspective. After the workshops these different perspectives can be analysed an compared against each other. Enabling a discussion of specific strength or weaknesses of the interventions, as well as informing designers an possible communication strategies in regards to these different fields of expertise.

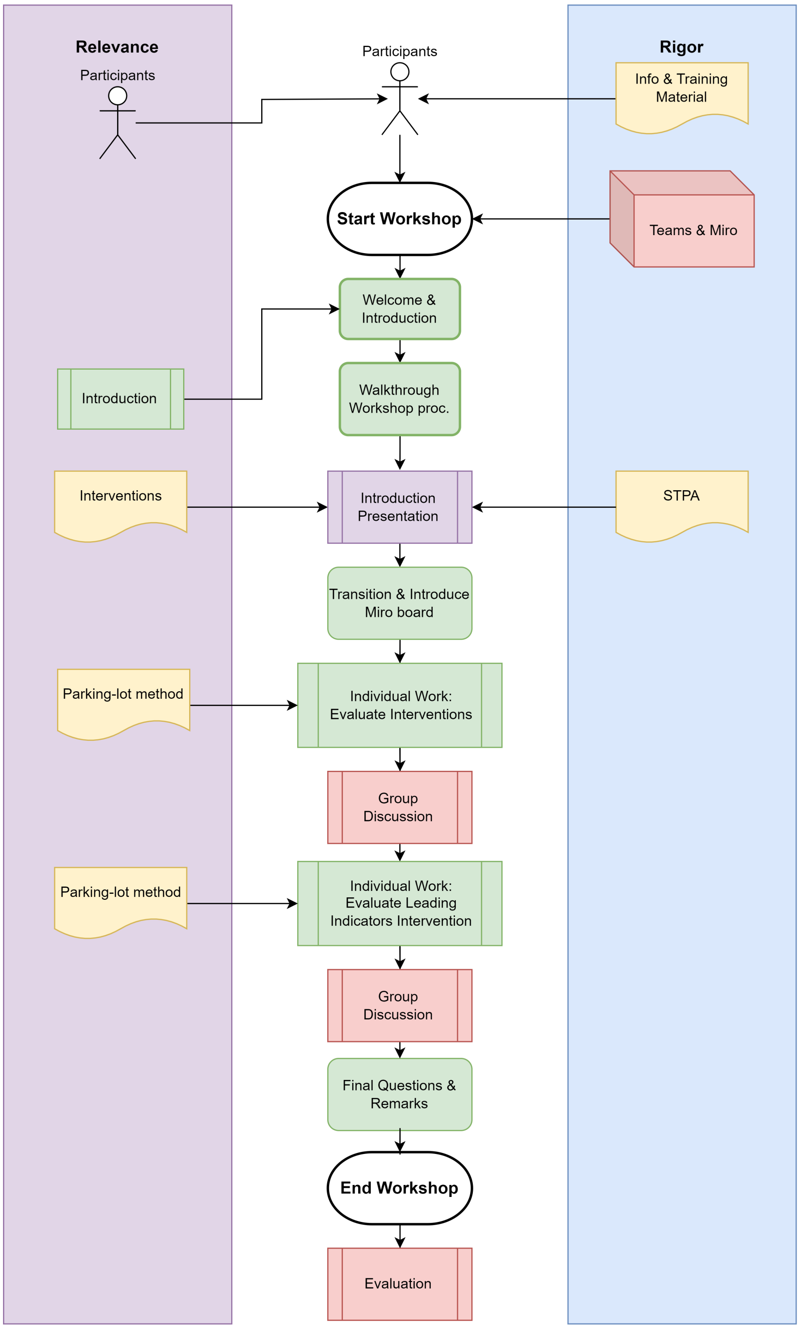


Figure 3.2: Workshop purposed process flow

### 3.3.2. Document Review

For the document review this thesis will utilize a variation of Retrieval-Augmented Generation (RAG) process. Making use of a large language model that is feed with contextualized information retrieved out of text segments from collected relevant documents. The process reviews official documents and news article reports detailing the events of the toeslagenaffair. The process is illustrated in Figure 3.3. The first objective is to reconstruct the existing safety control structure during the toeslagenaffair, while the second is to derive design objective recommendations from the literature surrounding the toeslagenaffair. During the process data is identified through official dutch government websites and archives. For the gathering documents several variations of search terms will be used the following queries give an impression of the utilized terms (translated from dutch):

1. ”( TITLE-ABS-KEY ( ”**toeslagenaffairl**” ) AND ALL ( **web** ) )” &
2. ”( TITLE-ABS-KEY ( ”**toeslagenaffair**” ) )” &
3.”( ALL ( **child benefit scandel**) AND ALL ( **netherlands** ) AND TITLE-ABS-KEY ( **web** ) AND ALL ( **emergent AND interaction** ) )”.

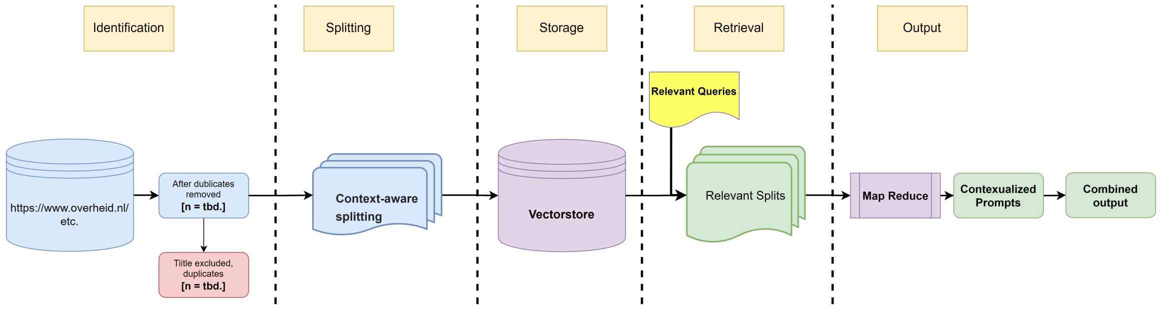


Figure 3.3: Document Review RAG Process

Further documents are collected through an iterative process searching through multiple documents on official website. As such backward and forward snowballing is utilized starting from core documents such as letters to the second chamber of dutch government.

Hereinafter the the documents are spit into context aware splits approximately 1000 token in length with an overlap of 100 tokens, using recursive splitting. These splits are then embedded with the help of the OpenAI API and subsequently stored in a vector database. The respective code can be viewed in the appendix B. Subsequently, relevant question regarding the context of the Toeslagenaffair are prompted to the vector database. These are then transformed into vectors themselves and a subsequent a similarity search is conducted to identify relevant text splits that closely relate to the posed question. The top 20 splits are then passed to the large language model (LLM) as context. Using a ’map reduce’ process these splits are finally summarized into one relevant context document, which the LLM uses to create an answer. Both the answer, as well as the original top 20 input splits are returned. The results can be viewed in appendix B. They build the bases for the information presented in Chapter 4.

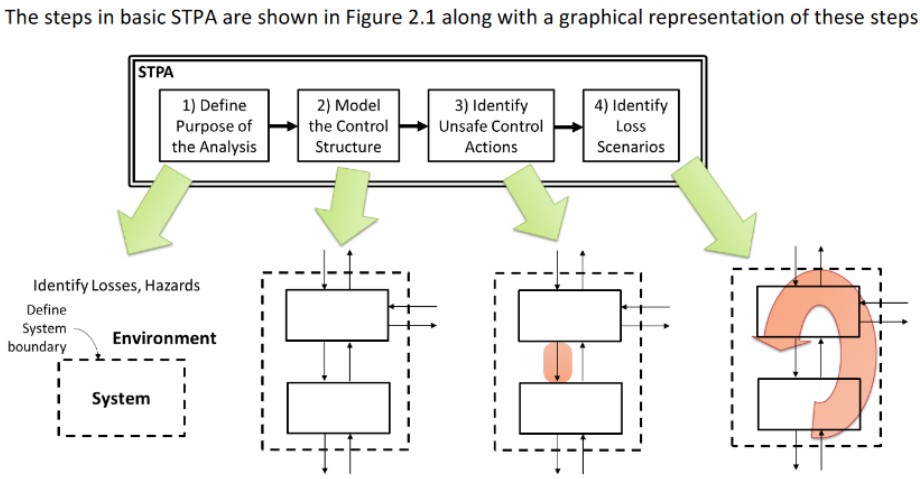
### 3.3.3. System Theoretic Process Analysis (STPA)

STPA is a hazard analysis and risk assessment technique used to identify potential hazards and vulnerabilities in complex systems (N. Leveson, 2011). It is a part of the System-Theoretic Accident Model and Processes (STAMP), which is a systems engineering framework designed to analyze accidents in complex sociotechnical systems. STPA focuses on understanding the system’s control structure and how it contributes to safety or risk. It emphasizes a system-centric view, looking at systemic issues rather than just component failures (N. Leveson, 2011). STPA is commonly used in industries with complex systems, such as aviation, healthcare, and nuclear power, yet can be applied to any complex system. It’s worth noting that the specific steps and details of an STPA analysis may vary depending on the context and the nature of the system being analyzed as such the process as depicted represents an idealization that is not strictly followed within this thesis.

Figure 3.4 illustrates the overall process flow of the STPA analysis derived from the STPA Handbook (N. Leveson, 2018). If differs in some ways to the outlined approach mentioned by Leveson in there book on system safety engineering (N. Leveson, 2011).

In the book Leveson outlines the process as follows(N. Leveson, 2011):

- *”1. Identify the potential for inadequate control of the system that could lead to a hazardous state. Hazardous states result from inadequate control or enforcement of the safety constraints, which can occur because:*
  *a. A control action required for safety is not provided or not followed.*
  *b. An unsafe control action is provided.*



The steps in basic STPA are shown in Figure 2.1 along with a graphical representation of these steps.

STPA
1) Define Purpose of the Analysis
2) Model the Control Structure
3) Identify Unsafe Control Actions
4) Identify Loss Scenarios
Identify Losses, Hazards
Define System boundary
Environment
System

Figure 3.4: STPA methodology according to STPA Handbook (N. Leveson, 2018)

*c. A potentially safe control action is provided too early or too late, that is, at the wrong time or in the wrong sequence.*

*d. A control action required for safety is stopped too soon or applied too long.*

- *2. Determine how each potentially hazardous control action identified in step 1 could occur.*

*a. For each unsafe control action, examine the parts of the control loop to see if they could cause it. Design controls and mitigation measures if they do not already exist or evaluate existing measures if the analysis is being performed on an existing design. For multiple controllers of the same component or safety constraint, identify conflicts and potential coordination problems.*

*b. Consider how the designed controls could degrade over time and build in protection, including*

*i. Management of change procedures to ensure safety constraints are enforced in planned changes.*

*ii. Performance audits where the assumptions underlying the hazard analysis are the preconditions for the operational audits and controls so that unplanned changes that violate the safety constraints can be detected.*

*iii. Accident and incident analysis to trace anomalies to the hazards and to the system design."*

This discrepancy may be attributed to the considerable time past between the two publications, allowing Leveson and others to refine and improve upon the STPA methodology. And secondly may be due to the differences in target audience each of the publications is directed towards. The process outlined in the STPA Handbook seems to follow a broader scope to make the analysis more approachable. It entails more general considerations that are a part of STAMP, but have relevance for STPA as well.

This research will utilize the flexibility of the method to make the outcome most suitable for the subsequent design task. As such the STPA analysis will be used to derive system constraints that can be used as requirements for the interventions that are subsequently designed.

### 3.3.4. CAST VS STPA

Hereinafter we will shortly outline our choice in favor of STPA befor CAST.

Causal Analysis based on STAMP, hereinafter called CAST, is another system safety method used to identify and understand the cause-and-effect relationships within a system. It aims to uncover the root causes of a problem or a set of problems. STAMP stands for System-Theoretic Accident Model and Processes. It is a framework designed to analyze accidents and other untoward events within complex systems. The goal of STAMP is to understand the systemic factors and organizational processes that contribute to accidents rather than focusing solely on individual component failures. Consequently, CAST facilitates the identification of essential questions necessary for understanding why an accident occurred (N. Leveson, 2011). Unlike traditional methods that pinpoint single causal factors, CAST allows for a holistic examination of the entire sociotechnical system design (N. Leveson, 2011). Leveson point out that this examination helps identify weaknesses in the existing safety control structure and suggests changes capable of addressing not only symptoms but potentially all causal factors, including systemic ones. Thereby CAST helps to shift the focus from assigning blame to understanding why accidents occur and preventing similar losses in the future. Achieving this goal involves minimizing hindsight bias and exploring the reasons behind people's actions based on the information available at the time (N. Leveson, 2011).

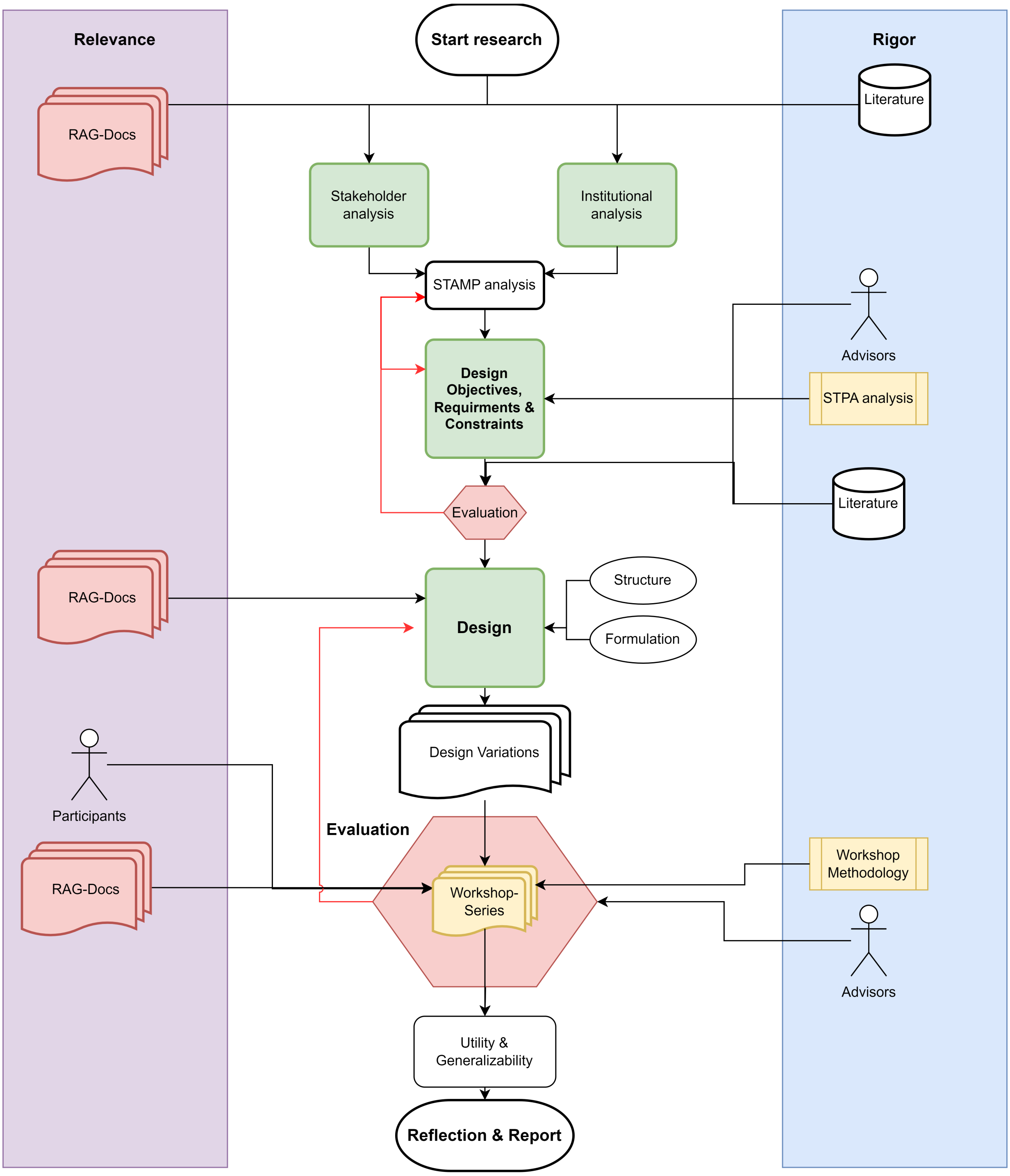
Based on this introduction one might suggest that utilizing CAST for this research represent a more suitable approach, because we are analysing a past event, during which an accident occurred. Based on the information provided, we can deduce that in such a scenario CAST would likely be more capable of identifying the ”root cause” for an accident in comparison to STPA. CAST would establish an event chain and analysis specific occurrences that happened during the affair in detail (N. G. Leveson, 2019). However we want to explicitly state that this is not the purpose of the analysis within this thesis. The Toeslagenaffair has been apply discussed in literature and public and the perpetuating system as bin adapted considerable. The purpose of this research is not to identify root causes for the Toeslagenaffair and improve the safety of the Toezicht process of the Belastingdienst through explicit recommendation. Rather we want to demonstrate the utility of system safety theory to improve system safety of complex social welfare systems that utilize ADM systems. As such the Toeslagenaffair only represents a means to an end, and the case was chosen specifically, because it has a known outcome (emergence of a digital cage) and fast body of documents discussing the systems state at the time. Because we want to create interventions that are applicable to other, similar systems and demonstrating the value of system safety is our overall objective, STPA is the more sensible choice. STPA provides more utility to practitioners because it is an ex ante not ex post analysis tool (N. Leveson, 2018). This means it can be used by practitioners to improve the safety of systems they are currently designing or operating. Not as with CAST, analyse the hazards in a system after an accident (emergence of cage) has occurred. STPA also takes a wider perspective on the hierarchical control structure and considers all possible unsafe control actions and their resulting system hazards not only once that have contribute to a specific incident, as with STPA. Ultimately, this results in a more detailed and complete system hazard list, which is used to derive design requirements (Hevner et al., 2004). Through this the utility and relevance of the design can be better assured.

## 3.4. Research Flow

The subsequent research flow diagrams depict the design and research process.

Figure 3.5 on the other hand illustrates the research process in it design context. It demonstrates what information sources are utilized in order to answer the different research tasks and thereby shows how ”relevance” and ”rigor” are induced into the research project.

Figure 3.5: Relevance and Rigor induced into Research Flow



# 4

# Socio-Technical System Analysis

This chapter lays out the socio-technical analysis of ADM system deployment and operation on the example of the dutch child care benefit scandal, known as the toeslagenaffair. It aims to answer the first sub-research question: *"What hazards in the socio-technical environment enabled the emergence of a digital cage during the ADM systems use in the dutch welfare administration at the time of the Toeslagenaffair?"*. The chapter initiates the analysis with a brief summary of the toeslagenaffair, before moving on to provide a more nuanced picture of the scandal relevant to the system safety analysis. Firstly, the stakeholders involved in the affair will be introduced, section 4.1. Secondly, institutional context of the toeslagenaffair will be explained, section 4.2. Both analysis are conducted to provide a reader with relevant background knowledge about the case and to outline the environmental conditions surrounding the system at the time. Hereinafter the analysis will focus on system safety control structure of the system that can be derived from the conducted document review, using RAG. Finally, a STPA analysis will be conducted and relevant hazards will be discussed, section 4.3. Throughout this chapter, relevant information stemming from the RAG document recovery methodology is used. The Results from this methodology can be viewed in Appendix B.

## 4.1. Dutch Child Care Benefits Scandal, Toeslagenaffair

The Dutch child care benefit scandal, also known as the "toeslagenaffaire," was a far-reaching scandal that involved the wrongful accusation of thousands of parents in the Netherlands of fraudulently claiming child care benefits. The scandal spanned several years, peaking in public and political attention around 2020 and 2021, leading to significant consequences for the affected families, as well as for Dutch politics and governance.

### 4.1.1. Origins and Development

The scandal centers on the administration of child care subsidies provided by the Dutch tax authorities, called "toeslagen". These subsidies were intended to support parents by partially covering the cost of child care, making it easier for them to work. Starting in the late 2000s, but intensifying around 2014, after the "Bulgarian fraud" case, where individuals exploited the Dutch welfare system to claim benefits unlawfully, the the Dutch tax authority (Belastingdienst) began flagging a large number of families as fraudulent claimants of child care benefits. Many of these accusations were based on minor errors or discrepancies in paperwork, or sometimes without any substantive evidence at all. However due to an opaque process structure within the Dutch Tax Administration and related organisation, these misclassifications were not detected. At the center of this development stood system first introduced by the Belastingdienst in mid 2013 called the 'Fraude Signalering Voorziening' (Fraud Signal System, FSV). The system was setup with in a the system landscape to detect and register signals related to potential fraud in the benefits system. The system was used to identify signals, register them, and investigate their validity and potential impact on the benefits system. The findings were then reported to relevant authorities for further action. However, concerns about privacy violations and misinterpretation of signals led to the deactivation of FSV in 2020. It later emerged that the tax authority had also used ethnic profiling in its efforts to identify potential fraud, targeting families with dual nationality or non-Dutch backgrounds. This aspect of the scandal highlighted systemic racism and discrimination within the government agency. The accused families faced severe financial hardship as they were ordered to repay the subsidies, often amounting to tens of thousands of euros. This led to debts, bankruptcies, unemployment, divorces, and significant mental health issues among affected individuals. In the worst cases parents lost custody of their children by court order, because they had fallen under the minimum standard of living. Many families fought for years to clear their names, facing a bureaucratic and legal system that was slow to recognize or rectify the mistakes. The government's aggressive pursuit of fraud, combined with a lack of effective avenues for appeal, exacerbated the situation.

### 4.1.2. Outcome and Aftermath

The issue received intermittent attention until investigative reporting and parliamentary inquiries brought it to the forefront. The government initially defended its actions but gradually acknowledged serious mistakes as evidence of wrongdoing and systemic failure accumulated. In January 2021, Prime Minister Mark Rutte's third cabinet resigned over their role in the scandal. This was a gesture to acknowledge the government's collective responsibility for the mishandling and misconduct. The Dutch government promised compensation to the affected families, including a standard payment of 30,000 euros to those unjustly accused of fraud. Additionally, there were calls for and efforts towards systemic reforms to prevent such failures in the future, including changes in the tax authority and the establishment of more robust checks and balances within the government. The scandal severely damaged public trust in the Dutch government and its institutions, raising questions about systemic bias, the balance of power, and the protection of citizens' rights. The situation attracted international criticism and concern, highlighting the potential for government policies and systems designed to prevent fraud to harm innocent individuals, especially those from marginalized communities. The full extent of its implications, both for the individuals directly affected and for Dutch society and governance, is still unfolding.

## 4.2. Stakeholder analysis

The purpose of this stakeholder analysis is to provide the reader with more context regarding the affair. This will shed more light onto the power dynamic that lead up to the scandal as well as the role actors played in resolving it. Putting the ensuing system safety analysis into this context allows for a better understanding of the environmental conditions relevant to the case.

### 4.2.1. Stakeholders

The toeslagenaffair involved several key entities, including the Dutch Tax authority known as ”Belastingdienst” and its administrative subsidiaries such as the ”Dierectie-Toeslagen”, ”Dierectie-Particulieren” (Individuals Directorate), ”Dierectie-MKB” (Small and Medium-sized Enterprises Directorate), the ”intensief toezichtteam” and the ”Landelijk Incasso Centrum (LIC)”. Several groups were also involved, such as the Combiteam Aanpak Facilitators, who were responsible for implementing enhanced supervision, and the Onderzoek effecten FSV Toeslagen, which investigated the effects of the FSV on toeslagen recipients. Aside from the Belastingdienst the victims registered in the FSV system, affected by the 'stopzetting' (termination) of toeslagen or subject to the unjust violation of their privacy. This also includes ”gastouderbureaus” (day care providers) as well as NGOs and individuals that supported and fought for the victims rights. Additionally, other governmental institutions such as the ”Raad van State” (Council of State), the Dutch Parliament as well as the Dutch Goverment (Rütte III) are involved in some capacity in the scandal. Several third party organisation impacted the affair in different capacities, such as the news organisations ”Trouw” and ”RTL News”, PwC and KPMG as an independent Auditor to the Belastingdienst, as well as third party developers such as Capgemini involved in the initial development of the FSV system. Finally, the Dutch public is also a relevant stakeholder whom's increased interest in the affair pressured may stakeholders to act in a transparent manner. The subsequent tables provide an overview of these different stakeholders their involvement, influence and motivation.

Table 4.1 shows the victims and related organisation and advocates. With in this section the role of the day care facilities as mediators between the parents and the Belastingdienst. Within this role they contributed inadvertently, in most cases, to victims sending incorrect benefit request to the Belastingdienst that subsequently lead to investigations. At the same time, these organisation were also targeted by the Belastingdienst and some cases wrongfully accused of fraud or negligence, where infact they were merely trying to support struggling families.

Table 4.1: Stakeholder Cluster: Victim & Advocates

| Stakeholder | Definition/Involvement | Roles & Influence | Motivations |
|---|---|---|---|
| Victims & Advocates | Victims and organisations as well as individuals related to the victim its inital and resulting circumstances | | |
| Victims & Families | Individuals and Families that were wrongfully accused of having commited welfare fraud or beeing negligent | - very limited influence to prove innocency<br>- facing administrative exclusion<br>- increasing strain on family ties etc. | - perceived and actual powerlessnes<br>- proving innocency<br>- pushed to the exitentail limit |
| Day Care Facilities | Day care facilities that were wrongfully accused of having commited welfare fraud or beeing negligent | - often provided victims with faulty advise<br>- where in the position to take advantage of leaniances in the dutch benefit policy<br>- could support struggeling families through this leaniance | - in most cases, support struggeling families<br>- sometimes, to receive higher subsidies |
| Advocates (NGOs & Lawyers) | NGOs and Lawyers that supported the struggle of the victims and day care facilities under scrutiny | Providing vitims with social and emotinal support, as well as fighting legal battle to profe injustice | - correct injustice<br>- help people in need |

Table 4.2 depicts the Stakeholder Cluster of the Belastingdienst. The Belastingdiesnt is the organisation from which the scandal was initiated. The different groups within the Belastingdienst had different roles within the creation of the scandal based on diverging motivations.

Table 4.2: Stakeholder Cluster: Belastingdienst and associates

| Stakeholder | Definition/Involvement | Roles & Influence | Motivations |
|---|---|---|---|
| Belastingdienst & Partners | Actors part of the Belastingdienst or related to it that were directly involved in the fraud detection process or utilized information originating from this process (third party government organisations) | | |
| Intensief Toezichtteam | The intensief toezichtteam was responsible for registering and analysing fraud signals in FSV and reponsible for forming decision on potential fraud cases | - take direct deicion on whether a citizen could be suspected of fraud<br>- take decision whether citizen benefits were incorrect<br>- take decision whether citizen commited fraud<br>- trigger ciminal investigations<br>- trigger benefit recollection | - Identify possible fraud cases<br>- Execute the interessts of the state and Belastingdienst<br>- Reduce feature fraud through consequent persecution of fraud |
| Dierectie MKB | MKB is responsible for small to medium sized companies, processing and handeling tax and benefit related tasks. | In relation to the toeslagenaffair:<br>- responsible for receiving and passing risk signals related to organisations or from organisation about individuals into FSV<br>- responsible for dealing with organisational fraud such as day care organisations supected of fraud | - Identify possible fraud cases<br>- Execute the interessts of the state and Belastingdienst<br>- Reduce feature fraud through consequent persecution of fraud<br>- coordinate with other departments about cases |
| Dierectie Particulieren | Departement is responsible for individuals, processing and handeling indicidual income statements and potential financial fraud related to it | In relation to the toeslagenaffair:<br>- responsible for receiving and passing risk signals related to individuals into FSV<br>- responsible for dealing with individual fraud such tax fraud | - Identify possible fraud cases<br>- Execute the interessts of the state and Belastingdienst<br>- Reduce feature fraud through consequent persecution of fraud<br>- coordinate with other departments about cases |
| Dierectie Toeslagen | Departement is responsible for the benefit program of the state, granting, payingout, withholding benefits as well as investigating fraud | - Directs the intensive toezichtteam, the FSV system and related groups in the organisation<br>- executes the policy of benefits on a holistic scale<br>- has overview of cross group developments | - Identify possible fraud cases<br>- Execute the interessts of the state and Belastingdienst<br>- Reduce feature fraud through consequent persecution of fraud<br>- coordinate with other departments about cases |
| Landelijk Incasso Centrum (LIC) | Departement within the belastingdienst responsible for collecting debt owed to the Tax authority, also handels incorrect benefits amounts and fraud cases | In relation to toeslagenaffair:<br>- Withhold benefits from individuals that were determined to have comitted fraud<br>- Demand payments or deduct owed amount from benefits if received benefit amounts were determined to be incorrect<br>- Reject individual payment plan request from citizens<br>- Enforce a full repayment of benefits if amounts were incorrect and fraud was suspected<br>- Collect full amount if fraud or neglegance was determined | - Enforce the regulatory guidelines regarding benefits<br>- Make sure the states receives its ’owed benefits’<br>- Consequent processing of fraudsters and individuals supected of fraud<br>- Ensure ”guilty or negligant individuals” are punished<br>- Maintain the states ability to collect owed debt with little leagal opposition<br>(obstruct information sharing to maintain advantage) |
| Ministry of Justice (prosecution) | Prosecutors responsible for bringing fraud cases to justice.<br>The criminal investigations would revolve around larger potential cases of fraud such as suspected fraud by day care faciliites | - Running criminal investigation<br>- Enabeling a prosecution of the organisations or individual who were determined to have comitted fraud<br>- Restricting access to Information for defendants | - Execute the interessts of the state and Belastingdienst<br>- Reduce feature fraud through consequent persecution of fraud<br>- Ensure frausters are punished<br>- Maintain the states ability to convict with little leagal opposition<br>(obstruct information sharing to maintain advantage) |

Table 4.3 depicts the other main governmental organisations involved in the scandal. These are organisations that initiated the scandal through legislation, policy and ruling decision, but also the organisations that resolved the scandal through the same actions. As an example, the cabinet under Rütte I as well as the Tweede Kamer during the period of the ”Bulgarian fraud” case were instrumental in demanding, reshaping and executing a stricter policy on welfare fraud investigation and punishment. Yet, through parliamentary inquires the members of the house of representatives also required the government to handle the affair transparently thereby enabling society to discover the full extent of the underlying administrative exclusion. So die Rütte III by taking collective responsibility, thereby ending the practice of cover up.

Table 4.3: Stakeholder Cluster: Other Govermental Organisations

| Stakeholder | Definition/Involvement | Roles & Influence | Motivations |
|---|---|---|---|
| Other Govermental Organisations | Stakeholders that played an important role in setting policy that lead up to the affair, and dealing with the repercutions after the scandel broke | | |
| Tweede Kamer | House of representative of the Netherlands | - pressured government to persuit welfare fraud after the "Bulgarian fraud"<br>- pressured government to transparently resolve the scandel report about wrongdoing came forth | - main tain checks and balances of the government<br>- pass legilsation that is in the interest of the Dutch public |
| Cabinet Rütte I | Government of the Netherlands, enacting legilation into policy | - reated to the "Bulgarian fraud" case to enforce stronger regulation conserning welfare fraud | - Execute the will of the people<br>- Remain in governance<br>- Implement party oppinions |
| Cabinet Rütte III | Government of the Netherlands, enacting legilation into policy | - initially enfored policy in line with Rütte I & II<br>- took political responsibility and stepped down | - Execute the will of the people<br>- Remain in governance<br>- Implement party oppinions |
| Raad van State | One of four high courts of the Netherlands, and counsile to the state | - Give advise on legilation and policy<br>- Make rulings regarding interpretation of law and policy<br>- Rulings allowed Belastingdiesnt to maintain a harshline on | - Interpret Law and policy to its best ability<br>- Maintain conistance in ruling<br>- Overrule injustice<br>- Act as balance |
| Landelijk Incasso Centrum (LIC) | Departement within the belastingdienst responsible for collecting debt owed to the Tax authority, also handels incorrect benefits amounts and fraud cases | In relation to toeslagenaffair:<br>- Withhold benefits from individuals that were determined to have comitted fraud<br>- Demand payments or deduct oweamount from benefits if received benefit amounts were determined to be incorrect<br>- Reject individual payment plan request from citizens<br>- Enforce a full repayment of benefits if amounts were incorrect and fraud was suspected<br>- Collect full amount if fraud or neglegance was determined | - Enforce the regulatory guidelines regarding benefits<br>- Make sure the states receives its 'owed benefits'<br>- Consequent processing of fraudsters and individuals supected of fraud<br>- Ensure "guilty or negligant individuals" are punished<br>- Maintain the states ability to collect owed debt with little leagal opposition<br>(obstruct information sharing to maintain advantage) |
| Ministry of Justice (juvenile law) | Justice departement that determines whether children could be placed out of their homes, due to abuse, financal instability etc. | In relation to toeslagenaffair:<br>- Determining whether parents could still care for their children based on the fincial and social facts relevant to the case<br>- Unaware of the benefit scandel etc. | - Determine the best possible way of living for the children in question<br>- Project children and juveniles |

Table 4.4 shows several important external third parties that influenced the affair through their involvement. PwC and KPMG played a pivotal role in the transparent resolution of the scandal by performing an transparent audit of the process within the Belastingdienst enabling a clear view of perpetuating factors within the organisation that aided the resulting digital cage to emerge. The news media, specifically, Trouw in cooperation with RTL News were instrumentally in bringing the scandal to the attention of the public and for front of the political debate subsequently enabling a transparent resolution of the scandal. Fraudsters are also a relevant party as their methods and ways of operation ultimately required the Belastingdienst to act. Their attempts of 'gaming' the social welfare system is underlying cause for increase scrutiny. Their nature of being highly 'adaptive' may have also played a role in Belastingdienst and Ministry of Justice ill-transparent communication strategy regarding potential cases of fraud. Inadvertently making it harder for wrongfully accused to clear their names. The third party developers involved in building the FSV system also impacted the outcome, by implementing software and IT infrastructure with what in hint side has proven to be meager understanding of the targeted system landscape and process. This can be revered back to the issue of incentive structure that will be discussed in the CAST analysis.

Table 4.4: Stakeholder Cluster: Relevant third Parties

| Stakeholder | Definition/Involvement | Roles & Influence | Motivations |
|---|---|---|---|
| Other thrid parties relevant | Thrid parties relevant to the creation or resolution of the scandal | | |
| Auditors (PwC & KPMG) | Auditors of the FSV system after scandal broke | - Provide an independent analysis of the situation leading up to the scandel<br>- invluence future public policy with report<br>- highlight issue with system under analysis | - generte revenue<br>- deliever on the pomisses to client |
| News Media (Trouw, RTL News) | Media involved in reporting the scandal | - swing public opinion<br>- investigate news stories<br>- highlight the issues groups of individuals are facing | - act as a social control mechnism<br>- stay relevant to earn revenue |
| Fraudsters | Organisations and indidivduals actually commiting welfare fraud | - gaming the system<br>- evoking the government to act<br>- utilizing weaknesses in the given structure, no matter the structure | - unethical opportunism |
| Third party developers | Organisations involved in developing FSV | - determined the system structure<br>- creating the data entrie and analysis protocolls<br>- creating access and integration protocolls | - generte revenue<br>- maximaiz revenue with limited resources |

### 4.2.2. Stakeholder progression

This sub-chapter illustrates the progression as well as activation pattern within the stakeholder clusters during the developments of the affair. As illustrated in Figure 4.1 the emergence of the "Bulgarian Fraud" cases and other related cases played a role in increase media and public interest in the topic of welfare fraud. This alleged the government specifically the House of Representatives of the Netherlands to act and pressure the Administration to take welfare fraud more serious. During this time cabinet members warned of the potential repercussions of a change in policy, but ultimately to no avail.

Legislation was amended (SCOURCE) and passed. The cabinet under Rütte I began enacting this legislation through several policy initiatives, mainly directed to improving oversight of the benefit grant process, surveillance of fraud cases and stricter prosecution and handling of individual suspected of fraud. These policies were enforced through the Belastingdienst and the Ministry of Justice. As part of enacting these policies third party developers were hired to develop and implement an improved fraud registration and surveillance infrastructure. Capgemini Nederland B.V. (Capgemini) was the organisation put in charge of this development. Capgemini was task with implementing the ”Dagboek Fraude Signalering Voorziening” (FSV) into the ”SOS .NET” which is an internal application framework of the Belastingdienst based on .NET framework developed by Microsoft. This system was then implemented into the existing system landscape in 2013 and started to be fully utilized during 2014.

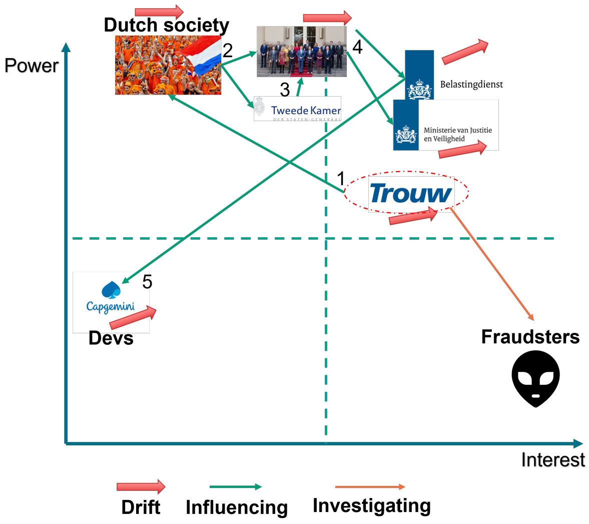


Figure 4.1: Power Interest Grid, in 2014

The policies, combined with newly implemented FSV application, lead to a strict processing on possible fraud cases. However due to a lack of an holistic system/process approach these citizens could be classified wrongfully, crucially without subsequent detection. This in turn lead to the ripple effects outlined in Chapter 1. Inadvertently a digital cage formed around the citizens victim to the administrative opaqueness, see Figure 4.2. Unfortunately, worsened by the fact the the Belastingdienst shared information from FSV with other organisations present in the ’RIEC’, namly the Police, the prosecution office (Openbaar Ministerie) as well as municipalities. Conversely, data in the FSV system could be used and interpret by these organisation to determine and coordinate controls. Further restricting Victims in their freedom and privacy rights. The harsh processing behaviour and invasive behaviour of the Belastingdienst and other administrative bodies then caught the attention of advocates and NGO groups that began to support the Victims social as well as legally. Ultimately, this resulted in the media gaining interest in these cases. Through the reports of the media the House of Representatives and the Cabinet began to realize that action needed to be taken and parliamentary inquires and committees were started. Simunantly, the ”Staat Raad” started to adopt another interpretation of the previously strictly enforce policy of ’full amount repayment’ after the background of the cases had been publicly analysed. Challenging the Tax authorities interpretation of the policy they enact. This resulted in the house and the Belastingdienst contracting independent auditors to survey and analyse the fraud signaling process of the Belastingdienst. Ultimately, uncovering the broad systemic administrative exclusion. Finally, this lead the cabinet of Rütte III to resign.

This progression that was followed within the Toeslagenaffair among stakeholders highlights the dual roles some key actors took in initiating and resolving the crisis. It can be deduced that the scandal, and hence the digital cage was in part triggered through a policy enactment resulting in a crack down on welfare fraud that casted the ’.Net’ to far. Aside from this, the incentive structure between the evolved parties can also be named as a perpetuation factor for the emergence of the digital cage.

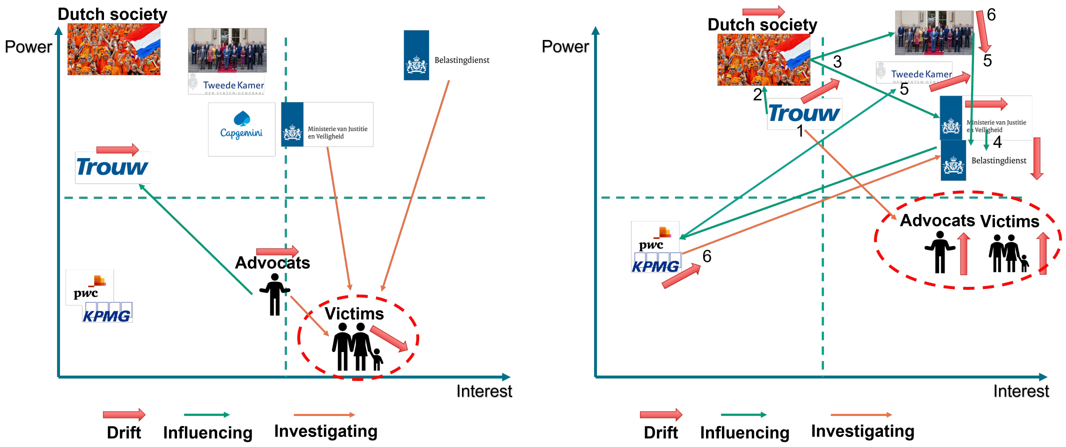


Figure 4.2: Power Interest Grid, between 2014-2020 & after 2020

Neither the Tax authority nor the third party developers involved in the creation of FSV had a clear incentive or motivation to make sure the ”wrong people were not harmed”, rather it seems, as several report state, that the focus was set on ”getting them”. Several of the FSV intrinsic design choices reflect this mindset. This analysis illustrates that aside from technical factors the stakeholder relations and interdependancies played a major role in how the ”Toeslagenaffair” unfolded. Subsequently, the institutional roles will be analyzed and described further.

## 4.3. Institutional analysis

ADM systems are the output of one or multiple government organisation to put policy into action. Consequently, the institutional within switch these systems are conceived need to be described. Subsequently, this thesis outlines the general process of putting policy into actionable governance, followed by the Dutch Government. Subsequently, institutional role of the Belastingdienst and several of its internal groups will be described. Here in after several important actors of the Toeslagenaffair will be grouped into their intitutional roles, to provide more context to their capabilities and roles. This analysis provided key insights into the larger control structure surrounding the toeslagenaffair. As well as enabling a better understanding of the environmental conditions relevant to the system safety analysis.

### 4.3.1. Process of policy in the Netherlands

The process of creating policy is central to any government. Democracies follow a rigorous process that evolves many different checks and balances. Policy can arise from different scenarios, yet usually they arise from situations that create a ”need for policy”. Such needs are then explored by inter-ministerial or parliamentary working group to determine the need for updates to existing legislation. Here in after that ”need for policy” or political intent is drafted into a Bill and proposed to the dutch Parliament. After the Bills have been reviewed, reveised and passed. A bill may ’assent’ after it has been determend to be lawfull. Such bills however represent high level guidelines and in practices these regulations then need to be implemented. Acts of Parliament often only address the main aspects of a topic. They provide for more detailed legislation in the form of implementation regulations, which - like Acts - contain generally binding regulations. They are not subject to approval by parliament. There are many kinds of implementation regulations. Some are required by law to be enacted by the government. These are called orders in council and take the form of Royal Decrees that must be signed by the King and one or more members of government. In other cases, a minister may be designated in an Act to enact more detailed rules. These take the form of a ministerial order. Finally, implementation regulations may drawn up by officials, if the Act in question allows for it. These regulations may provide for all sorts of powers, for example the power to issue licences or award grants. Often, there is some latitude as to how they will be put into practice. Ministers may lay down written rules about how to apply certain regulations. Sometimes these rules (contained in guidelines or circulars) are intended purely as instructions for civil servants, but in other cases they are published in order to keep the public informed.
In practice this results in interpretation space for Ministers and Ministries on how a specific policy may

be implemented. Those interpretation in turn can still have a large impact on the actual effects of a policy.

### 4.3.2. Belastingdienst

The Belastingdienst, or Tax and Customs Administration in the Netherlands, is entrusted with significant institutional power, primarily responsible for the implementation of taxation laws, the collection of taxes, and the administration of tax-related matters on behalf of the Dutch government. This power is deeply rooted in its mandate to ensure tax compliance, assess and collect taxes, and enforce tax laws, thereby playing a crucial role in the country's fiscal policy and management of public finances.

The authority wielded by the Belastingdienst is granted through a comprehensive legal framework established by Dutch legislation. The Dutch Parliament, consisting of the House of Representatives and the Senate, enacts tax laws and regulations. These laws outline the scope of taxes, rates, and the administrative procedures the Belastingdienst is obligated to follow. Furthermore, the Ministry of Finance oversees the Belastingdienst, with the Minister of Finance bearing political responsibility for the functioning of the tax system, including the policies and operations of the Belastingdienst (Belastingdienst, 2022).

In exercising its power, the Belastingdienst is authorized to collect various taxes, such as income tax, corporate tax, VAT, customs duties, and excise duties. It employs a range of tools for tax assessment, collection, and enforcement, including audits and assessments based on self-reported tax returns. Moreover, it can impose administrative penalties for non-compliance. Besides its role in tax collection, the Belastingdienst also provides services to taxpayers, offering guidance on tax laws, facilitating the tax filing process, and managing tax registrations and identifications. It also acts as main governmental facilitator for social benefits. In this function its the Belastingdienst responsibility to asses, determine and distribute benefits among the citizens in need of this support (Belastingdienst, 2022).

Several mechanisms are in place to ensure checks and balances on the power of the Belastingdienst. Citizens have the right to appeal decisions made by the Belastingdienst to independent courts, allowing for judicial review. The operations of the Belastingdienst are also subject to rules on transparency and accountability, necessitating the publication of reports and audits by the Dutch Court of Audit, which scrutinizes the government's income and expenditure. Additionally, the Dutch Parliament exercises oversight over the Belastingdienst's performance and policies, including through questions, debates, and the annual budgetary process, where resources allocated to the Belastingdienst are reviewed and approved.

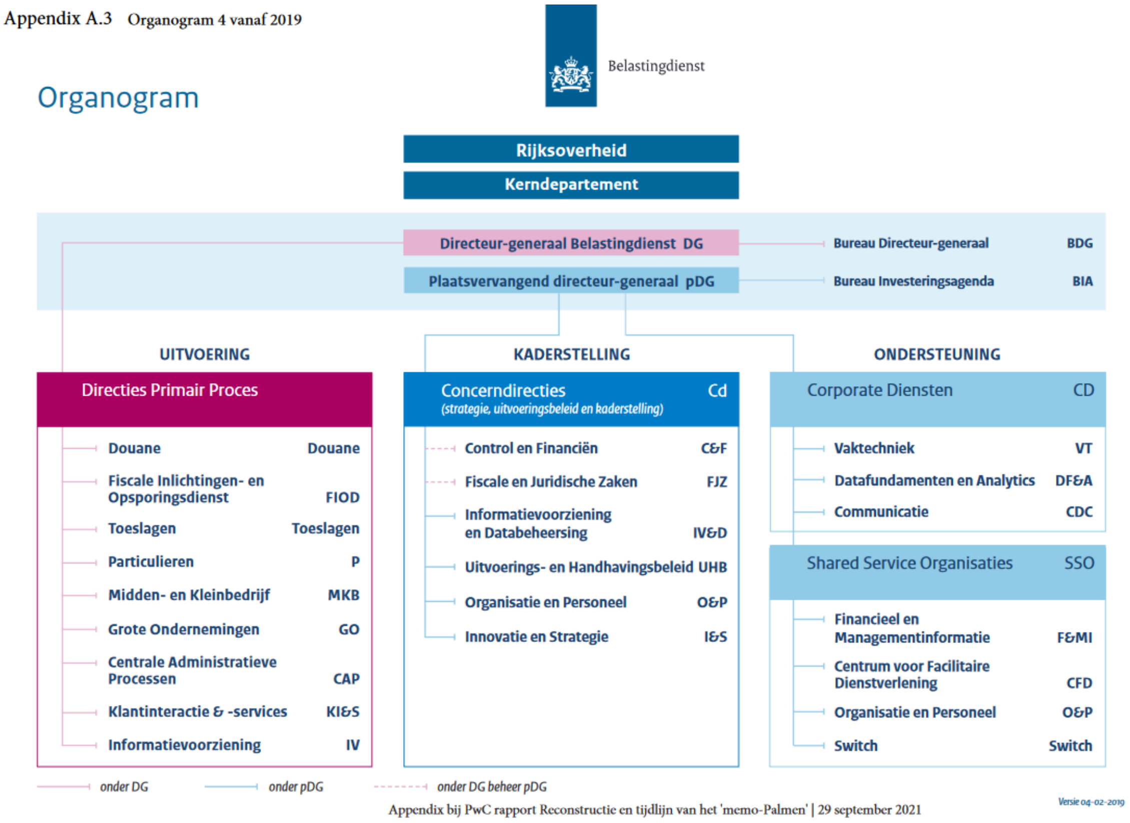
The Belastingdienst has several key mandates that are pivotal to the functioning of the dutch government system and society overall. In order to fulfill the mandates as best as possible the tax authority during the time of the toeslagenaffair was structured in the following manner (Persoonsgegevens, 2021):

Note that all the primary business processes are organized under the authority of the director general, while support organisations handling process and infrastructure are grouped under the authority of the vice director general of the Belastingdienst. This service structure allows the large sub-structures of the organisation to act quite autonomous, while utilizing personalized services, provided to them by the service support structure. Such a setup is necessary and makes sense since the primary process differ quite extensively in nature and regulation. This however also means that each primary process can be viewed as its own organisational silo. Inadvertently, making tasks that necessitate communication and cooperation between the different silos, challenging.

Subsequently major actors relevant to the Toeslagenaffair will be highlighted regarding their roles and responsibilities:

Toeslagen
”Toeslagen” translates to ”allowances” or ”benefits” in English. The key types of benefits they manage include:

- Huurtoeslag (Rent Allowance): This benefit is designed to help lower-income households afford their rent. Eligibility and the amount of aid depend on factors like income, rent cost, and the composition of the household.



Appendix A.3 Organogram 4 vanaf 2019

Belastingdienst

Organogram

Rijksoverheid

Kerndepartement

Directeur-generaal Belastingdienst DG

Plaatsvervangend directeur-generaal pDG

Bureau Directeur-generaal BDG

Bureau Investeringsagenda BIA

UITVOERING

Directies Primair Proces

| | |
|---|---|
| Douane | Douane |
| Fiscale Inlichtingen- en Opsporingsdienst | FIOD |
| Toeslagen | Toeslagen |
| Particulieren | P |
| Midden- en Kleinbedrijf | MKB |
| Grote Ondernemingen | GO |
| Centrale Administratieve Processen | CAP |
| Klantinteractie & -services | KI&S |
| Informatievoorziening | IV |

KADERSTELLING

Concerndirecties Cd

(strategie, uitvoeringsbeleid en kaderstelling)

| | |
|---|---|
| Control en Financiën | C&F |
| Fiscale en Juridische Zaken | FJZ |
| Informatievoorziening en Databeheersing | IV&D |
| Uitvoerings- en Handhavingsbeleid | UHB |
| Organisatie en Personeel | O&P |
| Innovatie en Strategie | I&S |

ONDERSTEUNING

Corporate Diensten CD

| | |
|---|---|
| Vaktechniek | VT |
| Datafundamenten en Analytics | DF&A |
| Communicatie | CDC |

Shared Service Organisaties SSO

| | |
|---|---|
| Financieel en Managementinformatie | F&MI |
| Centrum voor Facilitaire Dienstverlening | CFD |
| Organisatie en Personeel | O&P |
| Switch | Switch |

onder DG

onder pDG

onder DG beheer pDG

Versie 04-02-2019

Appendix bij PwC rapport Reconstructie en tijdlijn van het 'memo-Palmen' | 29 september 2021

Figure 4.3: Organisational Chart Belastingdienst from (Persoonsgegevens, 2021)

- Zorgtoeslag (Healthcare Allowance): Aimed at offsetting the cost of health insurance premiums for individuals and families with lower incomes. Like the rent allowance, eligibility and the allowance amount are income-dependent.
- Kindgebonden budget (Child Budget): A supplement for families with children, designed to help cover the costs associated with raising children. It is additional to the child benefit (kinderbijslag) and varies according to the number of children and the family income.
- Kinderopvangtoeslag (Childcare Allowance): This benefit supports parents by contributing towards the costs of childcare, making it easier for parents to work, study, or participate in reintegration activities. The amount of support depends on factors like the parents' income, the number of children, and the type of childcare. This is the allows most pressingly associated with the toeslagenaffair, and the scenario previously outlined in section 4.1.

The purpose of these allowances is to ensure that essential services like housing, healthcare, and childcare are accessible to those with limited financial resources, thus promoting social welfare and equality. With this mandate also comes the responsibility to check the benefits for their correctness, inspect and follow up on potential cases of benefit fraud.

MKB

The "Midden- en Kleinbedrijf" (MKB), which translates to "Small and Medium-sized Enterprises" division within the Dutch Tax Authority, Belastingdienst, is focused on catering to the tax affairs of small and medium-sized businesses in the Netherlands. This suborganisation plays a crucial role in the economic framework by supporting small business, which are fundamental to the Dutch economy. In essence, the MKB division acts as the bridge between the Dutch Tax Authority and the small and medium-sized business sector, facilitating a smooth interaction with tax systems and ensuring businesses can focus more on their operations while staying compliant with tax regulations. In this capacity the MKB is responsible to verify, correct, inspect and follow up on potential cases of fraud for this sector.

DF&A

The ”Datafundamenten & Analytics” (Data Foundations & Analytics) division within the Dutch Tax Authority, Belastingdienst, represents a modern approach to handling the vast amounts of data the authority collects and manages. This suborganisation is tasked with leveraging data science and analytics to improve tax collection processes, compliance, and overall efficiency. Here are the key functions and contributions of the DF&As division:

- Data Analysis and Insight Generation: Utilizing advanced data analytics techniques to analyze taxpayer and benefits receiver data, identify patterns, and generate insights that can help in making informed decisions.
- Supporting Tax Compliance and Enforcement: By analyzing data, the division can help identify instances of non-compliance or fraud more effectively. This allows for more targeted enforcement actions and supports the integrity of the tax system. To this purpose the decision runs several ”risk models” fitted to different ’use cases’.
- Policy Development and Evaluation: The insights gained from data analysis can inform tax policy development and evaluation. This includes understanding the impact of existing policies and forecasting the outcomes of proposed changes.
- Innovation and Technology Integration: The division is likely involved in integrating new technologies (like machine learning algorithms) into the Tax Authority’s processes. This could involve automating certain tasks, improving data management systems, or developing new tools for data analysis.
- Data Governance and Management: Ensuring the quality, security, and proper management of the data under the Tax Authority’s control. This is crucial for maintaining the trust of taxpayers and for the reliability of the analyses performed.

By focusing on data and analytics, the DF&A division plays a critical role in the modern Dutch Tax Authority, aiming to make it more efficient and effective to both the needs of taxpayers and the challenges of managing the nation’s finances (Belastingdienst, n.d.).

FIOD

The ”Fiscale Inlichtingen- en Opsporingsdienst” (FIOD), or the Fiscal Information and Investigation Service, is a critical suborganisation within the Dutch Tax Authority, Belastingdienst. Unlike other parts of the Tax Authority that focus on tax collection, administration, and compliance, the FIOD specializes in investigating financial crimes. Its mission is to combat fiscal fraud, money laundering, and other financial crimes that threaten the integrity of the tax system and the financial sector. The FIOD employs a range of specialists, including tax experts, accountants, legal experts, and investigators, who are adept at uncovering complex financial schemes. With advanced investigative techniques and technologies at its disposal, the FIOD plays a crucial role in enforcing the law, thereby ensuring fairness and integrity in the Dutch financial and tax systems.

CAP & LIC

The ”Centrale Administratieve Processen” (CAP), or Central Administrative Processes, is a suborganisation within the Dutch Tax Authority, Belastingdienst, that plays a pivotal role in the internal workings of the agency. It is primarily responsible for managing and executing the core administrative and logistical tasks that ensure the smooth operation of the tax authority’s various functions. These tasks include, but are not limited to, document management, data processing, and the overarching support of the tax collection process.

A significant subsidiary of the CAP is the ”Landelijke Incasso Centrale” (LIC), or the National Collection Agency. The LIC is integral to the Belastingdienst’s efforts in the collection of overdue taxes and fines. It operates as the enforcement arm for the collection of debts owed to the government, ensuring that outstanding amounts are paid in a timely and efficient manner. The LIC focuses its operations within the framework provided by the CAP:

- Debt Collection: The primary function of the LIC is to manage and execute the collection of debts owed to the government. This includes a wide array of debts, such as overdue taxes, fines, and other government-related fees. The LIC employs various methods to ensure these debts are collected, including sending reminders, making payment arrangements with debtors, and, if necessary, initiating legal proceedings.

- Enforcement Actions: If voluntary payment is not forthcoming, the LIC has the authority to take enforcement actions. This could include garnishing wages, levying bank accounts, or placing liens on properties to secure the payment of debts.
- Support and Guidance: While the LIC's main goal is to collect debts, it also provides support and guidance to debtors on how to manage their debts. If no malicious intent is detected, this includes setting up payment plans that take into account the debtor's ability to pay, thereby aiming to recover debts in a manner that is as fair and humane as possible.

The role of the LIC within the CAP highlights the importance of efficient administrative processes in tax collection and enforcement.

Recent challenges and calls for reforms have highlighted the need for improvements in the Belastingdienst's operations to enhance transparency, ensure fairness, and protect privacy. These developments underscore the importance of a balanced approach to the exercise of power by the Belastingdienst, ensuring that while it maintains the authority necessary for effective tax administration, there are adequate safeguards to check this power and protect taxpayer rights within the Dutch administrative and legal framework.

### 4.3.3. Raad van State

The ”Raad van State” (Council of State) is a key institution in the Netherlands, serving as an advisory body to the government and parliament on legislation and governance, as well as acting as one of the highest administrative court in the country. Its responsibilities and functions span various domains, reflecting its pivotal role in the Dutch legal and political system.

- Advisory Function

  Legislative Advice: The Council provides mandatory advice on proposed legislation and administrative orders before they are submitted to parliament.

  Policy Advice: Beyond specific legislative proposals, the government or parliament may consult the Council on broader policy issues.

- Judicial Function

  High Administrative Court: The Council of State acts as the highest court of appeal in administrative law, including disputes between citizens and administrative bodies. It reviews decisions made by lower administrative courts, government agencies, and other public authorities for their legality and adherence to principles of good governance.

  Dispute Resolution: In its capacity as an administrative court, the Council can annul or amend decisions made by governmental agencies, including the Belastingdienst, if they violate the law or principles of proper administration.

The Council of State's influence on the Belastingdienst primarily manifests through its judicial function. As the highest administrative court, the Council can review and judge the legality of the Belastingdienst's decisions. When taxpayers or benefit receivers appeal against decisions made by the Belastingdienst, these cases can eventually reach the Council if the initial appeals to the tax authority and subsequent legal challenges in lower courts are unsuccessful.

In such cases, the Council's rulings can directly impact the Belastingdienst by annuling cecisions, setting precedents, influencing organisational policy and administration.

Through these mechanisms, the Raad van State ensures that the Belastingdienst, like other government bodies, operates within the bounds of law and adheres to principles of fairness, legality, and good governance, thereby safeguarding citizens' rights and interests. In the case of the Toeslagenaffair, the previous rulings made by the Raad van State unfortunately perpetuate the administrators ability to harm citizens in specific cases. Among the questionable rulings, was the decision to allow the reclaim of 100% of the received benefits, if the amounts received did not check out with the eligibility of the individual in question (PwC, 2021).

### 4.3.4. Tweede Kamer

The ”Tweede Kamer,” or the House of Representatives, is the lower house of the Dutch parliament and plays a central role in the Netherlands’ legislative process. It is where most of the legislative work is done, including drafting, discussion, amendment, and voting on laws. The Tweede Kamer’s responsibilities and functions are broad and impactful, covering various aspects of governance and public policy.

The interaction between the Tweede Kamer and the Belastingdienst is a fundamental aspect of democratic governance, ensuring that the tax authority operates transparently, effectively, and in the public interest. Through its legislative, budgetary, and oversight roles, the Tweede Kamer has significant influence over the direction, efficiency, and accountability of the Belastingdienst.

### 4.3.5. Institutional Background

Below the previously described institutional and administrative organisations are allocated to their respective layer in Williamon’s layers for institutions. This helps the reader to gain an understanding of the different responsibilities each of the organisations fulfils and puts the different intitutions in relation to each other.

Furthermore this enables us to understand the environmental conditions surrounding the system at the time. We can better understand the influences that are exerted onto the system form the institutional side, through policy. The Belastingdienst in may ways can be viewed as an institution functioning on the third layer of Williamson’s schema. It can provide policy advise and through latitude in interpreting policy directives can also exert influence onto it directly. However its lower organisational functions are focused on enacting policy and finding ways to operationalize it. Since it are those organisational strucutres we will large be analysising, we must consider higher policy directive largly as given. Meaning they are ’environmental conditions’ surrounding our system of analysis. Policy eludes itself our direct influence during system safety analysis.

| Institutional actor | Williamson Layer |
|---|---|
| Tweete Kamer | Layer 2 |
| Raad van State | Layer 2 |
| Belastingdienst | Layer 2 & Layer 3 |
| Toeslagen | Layer 3 |
| MKB | Layer 3 |
| FIOD | Layer 3 |
| DA&F | Layer 3 |
| LIC | Layer 3 |

Table 4.5: Williamson Layers: Actors sorted (Williamson, 1998)

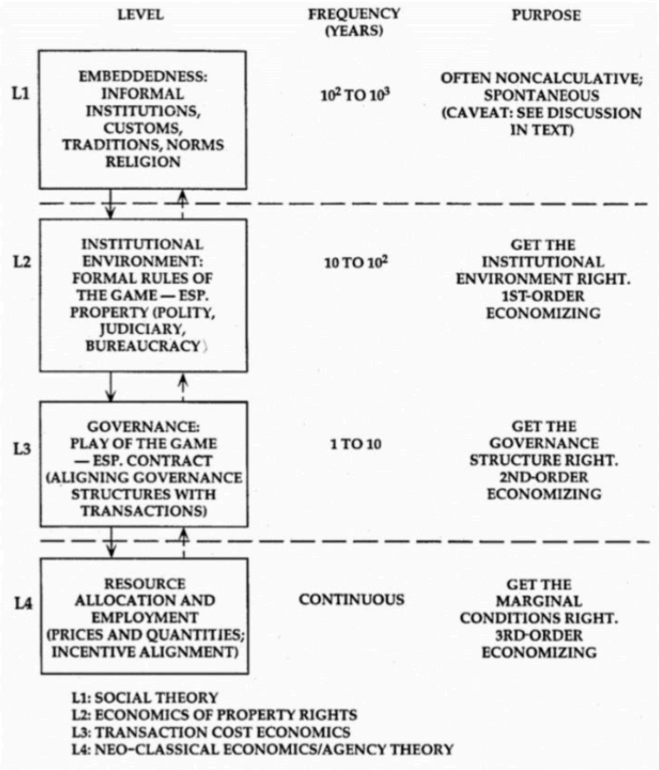

LEVEL
FREQUENCY (YEARS)
PURPOSE
L1
EMBEDDEDNESS: INFORMAL INSTITUTIONS, CUSTOMS, TRADITIONS, NORMS RELIGION
10² TO 10³
OFTEN NONCALCULATIVE; SPONTANEOUS (CAVEAT: SEE DISCUSSION IN TEXT)
L2
INSTITUTIONAL ENVIRONMENT: FORMAL RULES OF THE GAME — ESP. PROPERTY (POLITY, JUDICIARY, BUREAUCRACY)
10 TO 10²
GET THE INSTITUTIONAL ENVIRONMENT RIGHT. 1ST-ORDER ECONOMIZING
L3
GOVERNANCE: PLAY OF THE GAME — ESP. CONTRACT (ALIGNING GOVERNANCE STRUCTURES WITH TRANSACTIONS)
1 TO 10
GET THE GOVERNANCE STRUCTURE RIGHT. 2ND-ORDER ECONOMIZING
L4
RESOURCE ALLOCATION AND EMPLOYMENT (PRICES AND QUANTITIES; INCENTIVE ALIGNMENT)
CONTINUOUS
GET THE MARGINAL CONDITIONS RIGHT. 3RD-ORDER ECONOMIZING
L1: SOCIAL THEORY
L2: ECONOMICS OF PROPERTY RIGHTS
L3: TRANSACTION COST ECONOMICS
L4: NEO-CLASSICAL ECONOMICS/AGENCY THEORY

Figure 1: The four-layer model (Williamson, 1998)

Figure 4.4: Williamson Four Layer Model of Institutions (Williamson, 1998)

## 4.4. Systems Theoretic Process Analysis (STPA)

This Chapter builds the bases for the subsequent system safety analysis using STPA (Systems Theoretic Process Analysis), outlined in section 3.3.3. This chapter therefore follows the STAMP mythology outlined by Leveson (N. Leveson, 2011). STPA is usually used during the design phase of a new system or to improve an existing system (N. Leveson, 2018). On the other hand CAST, a method previously eluded to, was specifically designed to approach accidents that had occurred, in order to maximize learning from them (N. G. Leveson, 2019). Yet, as outlined in the methodology section this thesis utilizes STPA instead of CAST for two reasons. Firstly, the partial objective of this thesis is to demonstrate the utility of the system safety approach to ADM system deployment and operations in social welfare. In this regard, STPA represents the more use full method of analysis, because it provided more value to practitioner. Demonstrating the ability of STPA to improve system safety, will allow practitioners to draw direct inferences onto the system the may currently work on. Secondly, as it is the main objective of this thesis to design safety interventions that provide utility to other, feature ADM systems projects in social welfare, STPA provides a better bases for identifying the right design constraints. CAST may yield more insight into the specifics that lead to the Toeslagenaffair, yet post hoc analysis of the scandal have been done and amply discussed in Dutch public discourse, see (KPMG, 2020) also (PwC, 2021). Conversely, a STPA analysis fits better to the aims of this research.

The STPA process followed in this chapter will initiate with the discussion of the relevant system losses that were identified for the ”toeslagenaffair”. This aspect finalized the first part of the STPA ”Define Purpose”, which as previously already been initated through this research. It then moves on to the presentation toeslagen operating process at the center of the scandal, focused on the holistic process flow. This is done to provide the reader with a better understanding of the intricacies of the cases and to illustrate how the effects on citizens came to bear. This process diagram does not follow STAMP process depiction methodology. Control and feedback loops are omitted for better understanding. After, armed with a better understanding of the process, the hierarchical control structure of Toeslagen FSV system will be presented. This structure as well as the process outlined previously, will be used as a bases for the next step in STPA: ”Idenfity Unsafe Control Actions”. This step will be expanded to include feedback and data/information as well. Finally, the resulting system hazards and sub-hazards that result out of the unsafe control actions will be present, highlighted and discussed also touching upon the role of environmental conditions.

### 4.4.1. Losses

Losses represent something of value to the stakeholders of a system (N. Leveson, 2018). In this regard the losses determined for this analysis orient them self on the previously conducted stakeholder analysis. As this analysis is directed towards digital cages, the losses are based upon characteristics that constitute the digital cage, specifically L1 and L2. The respective unacceptable losses are however also well represented through the principals of good governance, since the system in question is a governance instrument, it should generally speaking, operate well if the good governance principals are meet. For a more better understanding and a more concise representation, these principals are however combined and contextualized into the following system losses:

| Loss ID | Loss Name | GG Principals |
|---|---|---|
| [L-1] | Mistreatment of Citizens/Organisations | 1, 5, 6, 9, 11 ,12 |
| [L-2] | Administrative Exclusion | 2, 4, 6, 9, 11, 12 |
| [L-3] | System Loss | 3, 7, 10, 12 |
| [L-4] | Waste of Resources | 3, 7, 8, 10, 12 |

Table 4.6: Losses relevant for STPA analysis

Within this loss list, no direct mention has been made of the need for the system to also identify fraud cases, however the ”system loss” in and of itself is integrated. This loss constitutes that the purpose for which the system was designed in the first case is being upheld through the system. At this point the thesis does not evaluate whether the systems purpose is correct, since this not part of the scope of the STPA analysis. Not worthy is therefore that the system should follow two objectives, prevent the emergence of a digital cage and justify its own existance through being an effective system.

### 4.4.2. Operating Process

The operating process of the FSV and its surrounding systems can be described as a number of interdependent processes. The control diagram in its entity can be inspected in appendix C, in this chapter sections of the diagram will be depicted to illustrate the described process. In general the process can be dissected into three main sub processes: (1) Information gathering, (2) Signal monitoring and decision making, and (3) Decision processing

Information gathering

The process conceived by the Belastingdienst starts with process of information gathering from various internal as well as external organisations. The due to its mandate as tax authority the Belastingdienst processes very large amounts of data trough various canals. However information relevant to the process of fraud detection for benefits are received and process mainly through the ”Fraud Meldpunt Toeslagen” (fraud desk benefits). The fraud desk benefits receives its information solely from internal sources namely the Infodesk FIOD, MKB Informatieloket (MKB infodesk) and from the risk models run by DF&A and DAT on big data provided by the CAP. All these three sub-organisations could receiv information either from external scources, such as regitration form from citizens or organisation or form internal sources redistributing information flow through the different silos. En example for that could be MKB receiving fraud signals from DF&A that are subsequently also forwarded to Toeslagen as well because individuals associated with that organisation may be receiving benefits. This was presumably the case for a number of the day care facilities related cases, previously mentioned. After the signals were received they were assessed towards their validity, registered in FSV and the BPM workflow management system. Hereinafter, depending of the priority the signal was classified with, either the Toezichteam line one or the Fraudteam Toeslagen would take over and asses, monitor and ultimately rule on the signal. Significant to mention is that at this stage signals were send to external third party government organisation without prior verification through either of the above mentioned teams by the Toeslagen Fraud desk.

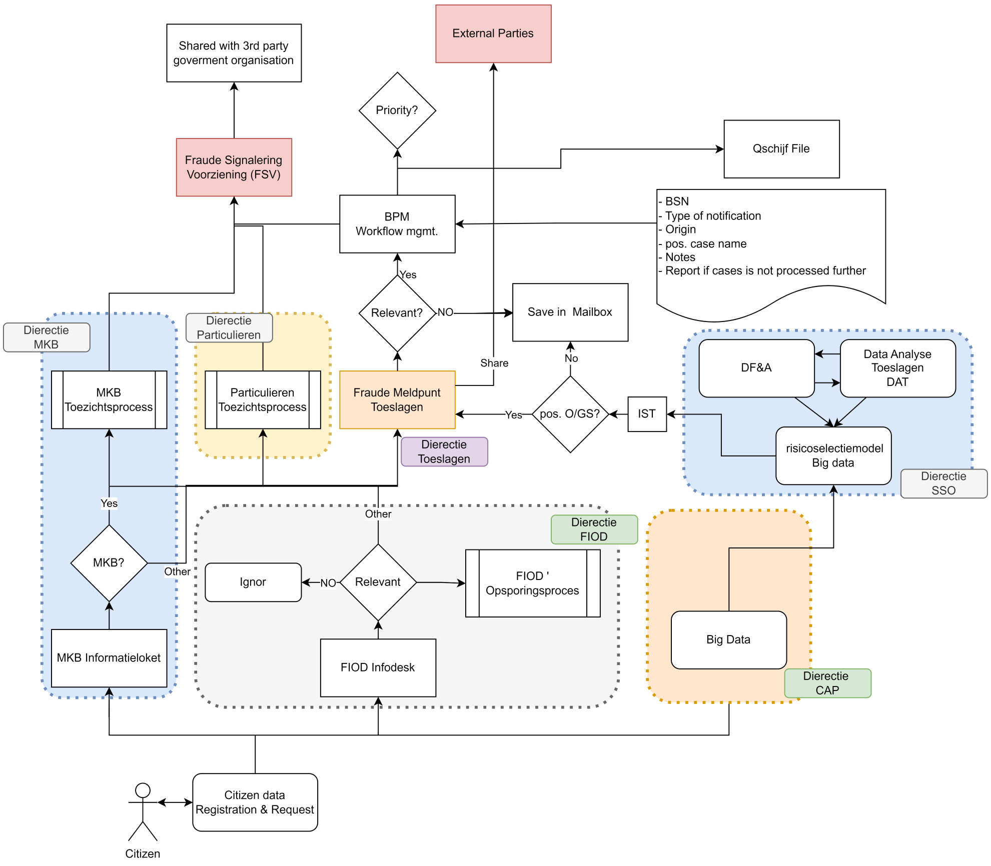


Figure 4.5: Data registration in FSV through the different organisational silos

Signal monitoring and decision making

The process of signal monitoring and decision making was either conducted by the 'Toezichtteam line one and two' or the Fraudteam, dependant on the priority and nature attributed to the FSV signal by the fraud desk Toeslagen. In either case the signals are accumulated for a specific individual receiving benefits. The teams then conduct analysis and make two key decision based upon the signals acumlated in the FSV system. Firstly, they determine whether the received benefit amount if correct ('Rechtmatig?') and if not whether they suspect an individual of fraud. If these requirements are meet this can already trigger the case workers to determine a "1x1" processing. Meaning any further request this individual makes has to be checked one on on by a case worker from the Fraudteam. Secondly, the case workers then determine whether the individual has committed fraud or acted with gross negligence. This decision is made by the Fraudteam or the Toezichtteams secound line. Depending on the decion the case is either forward to FIOD for criminal justice proceedings, in strong fraud cases. Or refereed to LIC, who are responsible of handling the incasso/recovery of the benefits from individual in question.

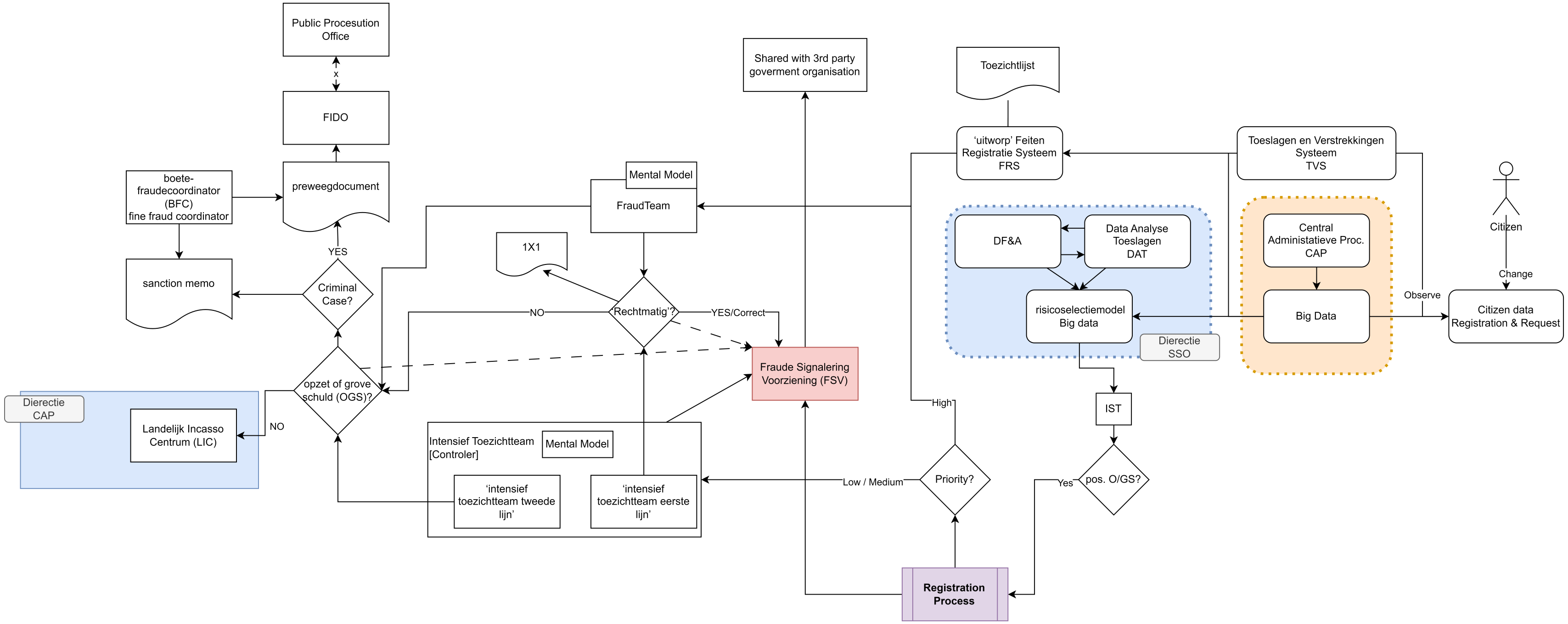

Public Procesution Office
x
FIDO
preweegdocument
boete-fraudecoordinator (BFC) fine fraud coordinator
sanction memo
YES
Criminal Case?
opzet of grove schuld (OGS)?
NO
Directie CAP
Landelijk Incasso Centrum (LIC)
Shared with 3rd party goverment organisation
Toezichtlijst
'uitworp' Feiten Registratie Systeem FRS
Toeslagen en Verstrekkingen Systeem TVS
Citizen
Change
Observe
Citizen data Registration & Request
Mental Model
FraudTeam
1X1
NO
Rechtmatig'?
YES/Correct
Fraude Signalering Voorziening (FSV)
DF&A
Data Analyse Toeslagen DAT
risicoselectiemodel Big data
Directie SSO
Central Administatieve Proc. CAP
Big Data
IST
High
pos. O/GS?
Yes
Priority?
Low / Medium
Intensief Toezichtteam [Controler]
Mental Model
'intensief toezichtteam tweede lijn'
'intensief toezichtteam eerste lijn'
Registration Process

Figure 4.6: Signal monitoring and decision making Toezicht and Fraudteam

Decision processing
During the final phase of executing the decision taken the case is either refereed to FIOD or LIC. The FIOD cases are in general referring to more high profile cases, the informational decision bases is often quit substantial. Consequences for the individual affect are more sever, with individuals facing criminal charges. Are only accesable on a case by case bases making them difficult to dissect in our process description. The majority of cases however were directed towards the LIC for debt collection of the individuals based on incorrect benefits received. Here it is important to note that the LIC received standardized information that included whether the individual was classified as a '1x1'. Crucially however the LIC determined this information to mean that Fraud had indeed been comitted and applied a number of stricter rules onto the individuals classified in this manner. Individuals classified as '1x1' were excluded from personal payment plans essentially necessitating them to pay back the 100% of the benefits received without debt resolution after 24 month. This put suffering individuals into a cage. Not that the LIC shared the information with other third parties and did not have a process of appeal in place for individuals that were classified as fraudsters.

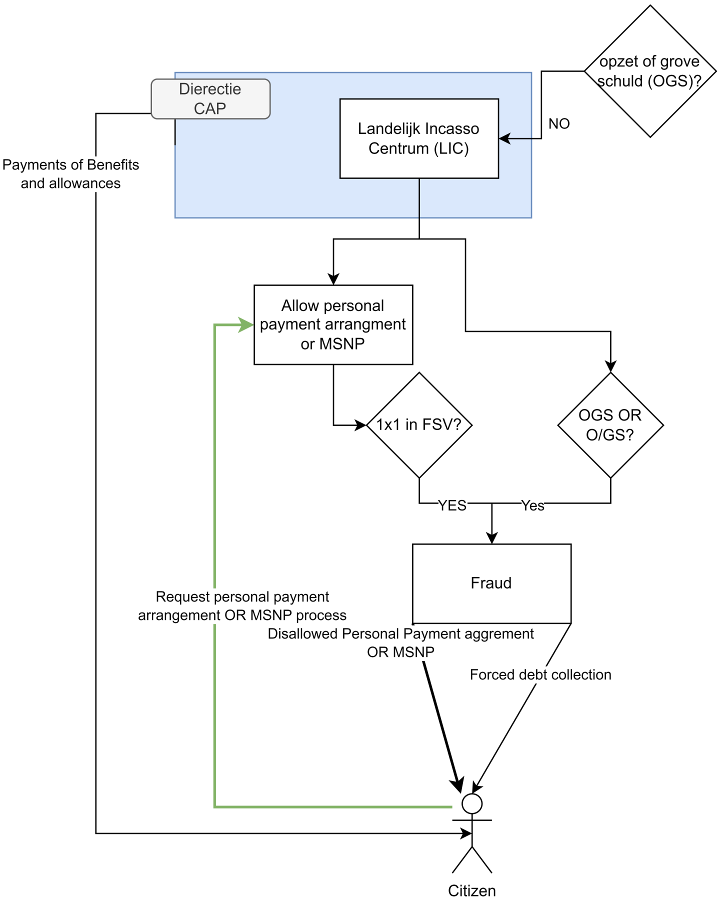

opzet of grove schuld (OGS)?
Directie CAP
NO
Landelijk Incasso Centrum (LIC)
Payments of Benefits and allowances
Allow personal payment arrangment or MSNP
1x1 in FSV?
OGS OR O/GS?
YES
Yes
Fraud
Request personal payment arrangement OR MSNP process
Disallowed Personal Payment aggrement OR MSNP
Forced debt collection
Citizen

Figure 4.7: Decision processing via LIC

### 4.4.3. Hierarchical Safety Control Structure

The entire hierarchical safety control structure can be viewed in appendix C. This section will only provide some general context to the control structure, by outlining the structural issues that could be identified.

As previously explained in the section 4.3.1 the Belastingdienst has several key mandates that are pivotal to the functioning of the dutch government system and society overall. As shown in figure 4.3 this results in each primary process of the Belastingdienst being its "own organisational silo". Inadvertently, making tasks that necessitate communication and cooperation between the different silos, such as fraud detection, challenging.

This Theme can also be identified when analysing the hierarchical safety control structure outlined in the appendix C. Here several observations grasp the views attention. Firstly, one can observe that the FSV system was utilized across a number of organisational silo. MKB, Particulieren, Grote Ondernemingen (GO), FIOD and Toeslagen each used FSV to register fraud signals. This also lead to a number of different registration processes, since each silo followed its own organisational directives. This also means that the signals registered in FSV are quite diverse in nature, since some apply to large organisations, some to small, and other to individuals.

Another observation forth noting, is that the organisation actually executing decision derived from the FSV system, the CAP more specifically LIC (for the case of Toeslagen), did not have direct access to the system but rather received output lists used as input to their systems.

Aside from the operational complexity that can be observed, one can also identify a complex structure during system development. Here there aspects are especially note worthy. Firstly, MKB was determined to be the business owner, a somewhat unusual decision considering that a number of organisation were appointed to utilize the system (KPMG, 2020). In a case were multiple organisations utilize a system it may have been sensitive to choose an impartial organisation as product owner. This will insure product requirements are negotiated openly and trade-offs are made from an impartial perspective. As pointed out previously, the silos organisational objectives and inner workings may differ quit substantially therefore they are likely to have and emphasise different requirements towards their respective systems. Secondly, the executing organisation was Generiek Kantoor Toezicht (GKT) an organisation that develops and runs applications for the primary process organisations, that at this point in time was organized in the silo GO (KPMG, 2020). Meaning another of the organisational users of the FSV system. It is unclear as to what degree this sub-organisation inherited the culture and ”working traditions” of its parent organisation GO, yet in general we can point out that determining a system developer that is link to one of the organisation using the system may inadvertently influence the development team to give preference to this use case. Such decision do not have to be conscious, they could be an expression/function of the developers past experiences made in this organisational environment. Thirdly, the actual system development seems to have been heavily influenced by a third party developer, namely Capgemini (Capgemini, 2013). This created a new dynamic in the development process that can not be ignored. While the utilization of expertise provided by external parties is common in the public sector and does provide a lot of benefits. There are also inherent risk that such decision invites into design and development process. The utilization of third party developers change the control structure within the organisation but also have an effect on the environmental conditions that impact systems.

Ultimately, the control structure makes apparent several structural flaws in the system that will be pointed out in the next subsection.

### 4.4.4. Unsafe Control Actions

The safety control structure as reviled several areas where the Belastingdienst system lacked adequate control, feedback or data/information. Feedback and data are causes for inadequate control and therefore are also included into this analysis.

Various types of inadequate control actions are identified, each contributing to systemic failures in Toeslagenaffair. They are listed in the appendix C. These inadequate control actions include cases where necessary controls are not implemented at all, referred to as ”Not Provided,” which results in the absence of mechanisms to detect and rectify errors or misclassifications. In other instances, controls may be ”Incorrectly Provided,” where the actions taken are inappropriate for the specific requirements of the situation, thus failing to effectively address the issue.

Another type of failure occurs when control actions are ”Too Late,” implemented after a significant delay that diminishes their effectiveness. This delay allows minor issues to escalate into more severe problems with broader consequences. Conversely, controls that are ”Stopped Too Soon” can lead to a recurrence of the initial problem if monitoring or corrective measures are halted before the issue is fully resolved. Additionally, controls that are ”Applied Too Long” can cause inefficiencies or unintended negative impacts by restricting system flexibility and responsiveness, which could stifle beneficial innovations or necessary adaptations.

The effectiveness of control systems in Automated Decision-Making (ADM) environments heavily relies on robust feedback mechanisms and the efficient flow of information. Inadequate feedback and information flow can significantly contribute to the perpetuation of control failures. Without robust feedback mechanisms, systems may not learn from past errors and continue operating under flawed assumptions, thus perpetuating a cycle of error. Inadequate or incorrect information flow can lead to misinformed decisions, where system controllers may overreact or underreact to situations. This issue is often exacerbated by siloed data systems that do not facilitate effective information sharing across different parts of an organization.

Furthermore, when feedback is delayed, it becomes challenging to respond promptly to changes or errors in the system. Delays may arise from slow data processing, the absence of real-time monitoring tools, or bureaucratic delays in communication channels. Even when feedback is available, it can be misinterpreted or ignored due to cognitive biases, a lack of expertise, or a misalignment between the system's outputs and the human operators' understanding. The following figure 4.8 taken from Levesons STPA handbook (N. Leveson, 2018) illustrates the above delineated issues.

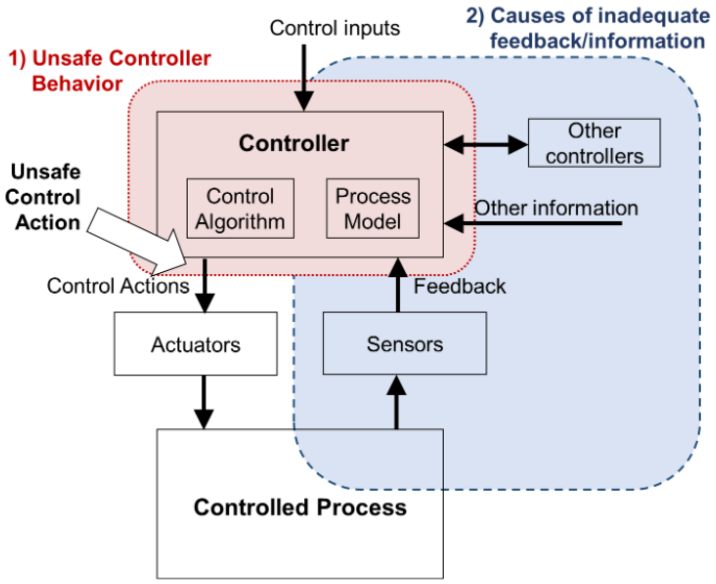


Figure 4.8: Causes unsafe Control Actions(N. Leveson, 2018)

### 4.4.5. System Hazards & Sub-hazards

After having outlined the hierarchical control structure of the system/process in question we now turn towards the system hazards and sub-hazards that are identify able trough the safety control structure, provided the losses that have been outlined before hand.

The STPA handbook defines a hazards as: "[...] a system state or set of conditions that, together with a particular set of worst-case environmental conditions, will lead to a loss." (N. Leveson, 2018). The guide outlines three rules by which to define hazards (N. Leveson, 2018):

- Hazards are system states or conditions (not component level causes or environmental states)
- Hazards will lead to a loss in some worst-case environment
- Hazards must describe states or conditions to be prevented

In accordance with these guidelines the following hazards have been identified:

The above outlined hazards represent a clear illustration of the systems hazardous state that lead up to the toeslagenaffair. They highlight nicely, that several unsafe system conditions that lead to the toeslagenaffair. Additionally they show that in reality it is hardly possible to single out and blame individual groups or parties. They hence underline Leveson's message nicely: Safety is a structural issue (N. Leveson, 2011).

To build inferences to the previously discussed theory in this thesis, subsequently, some of the key hazards will be discussed in the context of systems theory, more specifically adaptation and emergence.

| ID | Hazard | Losses |
|---|---|---|
| **[SH-1]** | **Process & Model misclassifies citizens** | **L1** |
| Organisation | | |
| [SH-1.1] | Insufficient, opaque and slow process comunication between subsystems | |
| [SH-1.2] | Information from citizens is shared prior to signal verification | |
| [SH-1.3] | Insufficient holistic process measuarbility and visability | |
| [SH-1.4] | Diverging Mental models of process | |
| [SH-1.5] | Diverging data registration process for differen organisations silos regarding FSV system | |
| Data | | |
| [SH-1.6] | Insufficient data quality in data registration systems | |
| [SH-1.7] | Bureaucatic, pungent/strict, & suspicious culture towards subjects | |
| [SH-1.8] | Designd data structure in data registration systems are inadequate | |
| Model | | |
| [SH-1.9] | Opaque Risk model structures | |
| [SH-1.10] | No logging functions tracking decisions (For risk models and FSV) | |
| [SH-1.11] | Different understanding of model objectives between stakeholders DAT (Accuracy), Toeslagen (Specificity) | |
| [SH-1.12] | Lack of model causality | |
| [SH-1.13] | Unfairnes and Biases in ML/AI pipline | |
| **[SH-2]** | **No citizen support process** | **L1, L2** |
| [SH-2.1] | Bureaucatic, pungent/strict, & suspicious culture towards subjects | |
| [SH-2.2] | No logging functions tracking decision making | |
| **[SH-3]** | **Data protection standards for third party information sharing are low** | **L1, L2** |
| [SH-3.1] | Toeslagen shares data to thrid party government organisation, pior to signal verification | |
| [SH-3.2] | Insufficient data quality in data registration systems | |
| **[SH-4]** | **No Recovery protocolls for process/system failure in place** | **L1, L3, L4** |
| [SH-4.1] | Ridgied process and system structure | |
| [SH-4.2] | Insufficient, opaque and slow process comunication and visability between subsystems | |
| [SH-4.3] | MKB is buisness owner of the FSV system | |
| **[SH-5]** | **Insufficent system correction/improvment** | **L3, L4** |
| [SH-5.1] | Insufficient, slow and opaque feedback to development | |
| [SH-5.2] | Insufficient development resources for system correction and improvement | |
| [SH-5.3] | Long change request/feedback process | |
| [SH-5.4] | MKB is buisness owner of the FSV system | |
| **[SH-6]** | **Operators are unaware of current system state in realtion to safety** | **L1, L2, L3** |
| [SH-6.1] | Key process stages are missing feedback loops from down stream work organisations | |
| [SH-6.2] | Control actions to measure system state are missing at Key process stages | |
| [SH-6.3] | Reports about system state are miss interpreted | |

Table 4.7: System Hazards present in Toeslagen FSV system

From the system hazards we gather that several core aspects that were missing within the FSV system and the surrounding system landscape that are pivotal to system safety. Firstly, several of the described hazards point towards a clear lack in ability to measure, interpret and hence ultimately know the system state, in relation to safety.

Secondly, we can infer that the system designers made several, key assumptions about the system and its environment during design that ended up being wrong. With no means of measuring the system state, these assumptions could not be verified or clearly identified as being broken.

Thirdly, the systems hazards clearly show a lack of holistic process control directed towards managing the system as a hole. This in turn uncovers the absence of a shared ”mental model”. A coherent mental model, as Dobbe et al. (R. Dobbe et al., 2021) outline, is a fundamental prerequisite for safe automated decision making systems design. End several of the resulting harm can directly be attributed to control actions that were unsafe due to the underlying discrepancies between the mental models of to sub-organisation / systems.

Another key system hazard is the lack of a clear data ontology both in relation to incoming, processed signals in FSV and outgoing signals to third party organisations. This lack of data ontology, contributed to a perpetuation of the misalignment of mental models, and overtime had to affect that the data quality of the system was allowed to deteriorate. It hence had a clear impact on the misclassification that resulted. Furthermore, it directly contributed to the emergence of the digital cage, by allowing other organisations to interpret the information they received from the Belastingdienst in their own ways. These would then converge onto citizens and impose unnecessary constraints and harms onto them.

A perpetuating hazard, as also the clear lack of sufficient feedback between operators, and managers as well as system developers. This pared with the overall slow and opaque process for system improvement through change requests, meant that the system could not adapt to changing circumstances and environmental conditions fast enough. Resulting in a rigid structure that neither of the system user organisations could be entirely happy with.

The aforementioned hazard was strengthened by the clear technical focus exhibited by product owners and designers. This meant that larger organisational relations or emergent effects were left unde-

tected from the start. Ultimately, this exemplifies the insufficient safety culture present in the system at the time. Appendix D underlines this reasoning, it shows is the use cases specification provided by Capgemini in October 2013 for the FSV system. We can gather that the Use cases are very technically focused, and analysis of the cases specified also illustrates a clear lack of inherent control actions incorporated into the system in the form of system and data safety, quality performance functions.
Finally, the system was not equipped with sufficient recovery and mitigation actions, for when issue with the system may occur. This further exemplifies the short mindedness exhibited by the system designers, and ultimately proved to be instrumental in the harms that the system inflicted upon citizens. Several actors had raised concerns about the systems operational state, well prior to the uncovering of the scandal. With a system recovery plan in place it is likely, that a system review would have been triggered earlier, making the mitigation of some of the inflicted harm likely. However since such a process was omitted, the actions of decision makers was to ”avert and ignore”, since nobody wanted to be held responsible.

### 4.4.6. Environmental conditions

These hazards combined with environmental worst case conditions will cause a loss. The environmental condition can vary in this respect. In the case of the Toeslagenaffair, however several environmental conditions were instrumental in causing the resulting digital cage. These will briefly be summarized here, to illustrate the applicability of the hazards, as well as to highlight the environmental conditions, since the are applicable to other, similar cases.

Policy
Policy is a critical environmental condition to any operating system in social welfare administration. At the same time policy represents a higher level system component that imposes constraints onto the system and its behaviour. However, for system designers primarily concerned with the safety of sub systems in social welfare, policy often represents an unchangeable condition and can has be understood as an environmental condition effecting the system, since their is no influence that can be exerted directly offer this condition, see (N. Leveson, 2011). As described in section 4.3 policy is critical, as it can influence the executing administrative organisation on several levels. Policy dictates broad and specific operational regulations that the organisation must adhere to. This sets specific constraints for the organisation to adhere to. Overall these constraints create a safer system structure, but on the other hand policy that perpetuates unsafe system states exist and they majorly contribute to harms arising. A wide spread example for such a policy is for example the pressure to cut back on budgets. Budgetary pressure can get in the ways of safety by representing a conflict of interest, resulting in a trade-off that jeopardizes the system safety. It is unclear whether is aspect, specifically, played a role in the formation of the toeslagenaffair, but budgetary constraints could provide a possible explanation as to why one specific client organisation was chosen as product owner instead of an overarching, independent organisation such as the ’Informatievoorziening’ or a committee consisting out of all product clients. As discussed in chapter 1, ADM system themselves can be a product of the need to improve ”efficiency”. Ultimately, such converging system goals lead to hazards and harms if safety is viewed as a tradeable characteristic of the system or simply not considered as a metric that might be impacted negatively through such decisions. Such trade-offs need to be consciously considered during the formation of policy.

Additionally the laws that are enforced through policy can evoke direct harms, since they specify punishment and incentives. If such laws are not devised well, or interpreted in a stern manner, they together with system hazards form harms. In the Toeslagenaffair, the example that can be used is the decision to reclaim 100% of benefits once a disparity between the amount that was received and the amount that an individual was eligible to had been detected. This decision that was grounded and confirmed by dutch law at the time pared with the system hazard of misclassification, ultimately lead to harm. Hence, policy and its underlying laws directly enable existing system hazards to be perpetuated onto citizens inform of real world consequences.

Crucially also, policy evokes culture. Since policy acts as a system constraint and directive the nature of policy is often adopted by the executing administrative bodies. Inadvertently, resulting in organisations that, depending on the policy directive, focus on performance and effectiveness rather then safety. Such cultural characteristics are especially problematic in social welfare administration, since the decisions that are made are sensitive in nature and often impact more vulnerable groups of society, as Alston previously eluded to (Alston, 2019). Through policy politicians give critical impulses to the institutions in charge of executing them. Hence to prevent digital cage it is also vitally important to critically asses the underlying cultural implications of policy. The toeslagen case is a prime example for this as the recent report of the tweede kammer on the role of Parliament during the affair outlines. A fitting manifestation of politics being aware of this relation is the statement former secretary of state for finances Frans Weekers made during the parliamentary debate about stricter benefit fraud persecution after the ”Bulgarin fraud” case in 2014. He is said to have said: ”de goeden onder de kwaden zullen lijden” (the good ones among the bad ones shall suffer) prior to bowing to the parliamentary pressure to tighten welfare benefit control(“Toeslagenaffaire”, 2024). Ultimately, this policy lead to a ”get them culture” in the Belastingdienst that shifted the focus from safety of decisions to prosecution of potential fraudsters (Persoonsgegevens, 2021).

Finally, policy also creates issues for system development. Policy changes that are enacted by Parliament often constitute an extensive change in direction and adaption of law and administration. This inadvertently leads to the fact that the systems used to enforce these policies are large and often loaded with numerous features to represent all the intricate parts and ruling changes, enacted through the policy. This represents a natural conflict between policy that should be as specific as possible and information systems that should be keep as simple as possible. Introducing a large number of different features in one system deployment immensely increases complexity and is often cited as the number one reason why system deployment fail (Kim et al., 2016b). Same can be assumed for the Toeslagenaffair, appendix D exemplifies the extensiveness of the use case size on the example of the FSV system. In total the system was to service 225 use cases, with varying complexity in one web application (see appendix D, ”2.1 Use cases”, Page2 & ”2.2 Applicatieonderdelen”). It is unclear in how many deployment cycles these different use cases were deployed, this is not further specified. Yet the full document does also not specify or refer to an extensive deployment plan, suggesting the system was implemented in rather large batches (Capgemini, 2013).

Workload

Workload refers to the amount and complexity of tasks that the system is required to handle within a specific time frame. This encompasses the volume of data the system processes, the number of decisions it needs to make, the speed at which these decisions are required, and the complexity of the sub-processes involved in making these decisions. One can clearly see how this environmental condition can have an impact on the overall decision outcome of a system.

An increase in workload can lead to system performance degradation. As the system tries to process more information than it is designed for, it may slow down, leading to delayed decision-making or, in severe cases, temporary system failures. This can be critical for ADM systems used in safety-critical applications, such as for fraud detection. More critical however is the fact that an increased workload, also increases the likelihood of errors in decision-making. ADM systems, particularly those employing complex algorithms and machine learning models, may produce less accurate or reliable outputs under stress. This could be due to overfitting, where the system becomes too tailored to the specific data it has seen and fails to generalize well to new, unseen data under high workload conditions. A high workload also impact the case workers that follow and control the advise provided by ADM systems. Under higher stress they are more likely to concur and confirm the systems results, rather then to question and investigate each case in detail. As the system becomes overloaded, the quality of the decisions made can suffer. This might happen because the system may need to cut corners to keep up with the demand, potentially leading to unsafe outcomes. This also applies to the data that downstream work centres are feed with. If for example an unusual number of cases had to registered in the FSV system in the Belastingdienst, this could have a significant effect on the data quality registered in the system. This could evoke a “trash in trash” out principal, something that also aligns with the findings made by PwC during their report on the scandal. PwC found the FSV system to by a case of “roten infrastrucute” (PwC, 2021). This suggest that the system itself suffered from a degradation of data over time. A consistently high workload can also make it difficult to perform regular maintenance and updates or change request on the system. This can lead to the system operating with outdated algorithms or decision-making criteria, processes, further impacting its safety.

Causality in ML and AI
Another important technical factor to consider is the relation of causality within machine learning and artificial intelligence applications. Machine learning models are exceptionally good at identifying patterns and correlations in data. However, correlation does not imply causation (Lin & Ikram, 2020). This distinction is crucial because acting on correlations without understanding causality can lead to potential harms further down the road. This issue of causality also means that often operator will face difficulties in explain the models behaviour and reasoning. This aspect can be viewed as an environmental condition, because it relates to the natural characteristics of ML and AI and hence will always impact any ADM system. It does not however have to translate directly to a system hazard, since if this aspect is accounted for within the system design its impact can be mitigated. If not addressed this environmental condition can easily combine with existing system hazards to create harm. If for example the ADM models implemented are not version controlled and decision outputs are not regularly verified. We need to version control our model because each model represents its own instance, trained on different data, but because we models acts upon patterns not causality, model stability remains fragile (Breck et al., 2017). Issues such as data or concept drift can not be extensively discussed here, but act as strengthening arguments in this case (Côté et al., 2023).

Gaming
It is reasonably to assume that governmental policy is often if not always subject to gaming, where ”gaming” in this context means stakeholders (individuals, corporations, interest groups, etc.) attempt to influence, exploit, or maneuver within the policy framework to achieve beneficial outcomes for them. Gaming does not inherently imply bad intentions. The motivations and impacts of gaming can range widely from benign or beneficial to manipulative or detrimental, depending on the context and the perspectives of those involved. The Toeslagenaffair is a fitting example for this. Several groups exerted their influence on the system or tried to ’Game’ it. The most obvious group are actual fraudsters, that through their actions first evoked the believes in officials to have to crack down on fraud in the welfare state (PwC, 2021). This groups follows inherently immoral intentions that are directed towards exploitation of the welfare state. From a system perspective it is very difficult to adjust to this group, since the fraudsters are actively trying to deceive the system, they are capable of ’adaptation’ and will change their methods and approach once they suspect that their previous tactics are no longer successful. Their general approach is to ’mask’ their behaviour to make it look similar to that of rightful benefit receivers. This in itself creates a risk for system operators that are looking to utilize ADM tools. These are based on historical data that might not prove to be accurate for current tactics of fraudsters any longer. Furthermore, both overfitting as well as generalization could potentially lead to misclassification of citizens. If the models are keep to general, a large number of possible ’false positives’ may be created. While if the models are ”overfitted” they are lightly to focus to much on very specific features, overemphasizing the importance of this specific ’discriminator’. This can result in the model exhibiting to much confidence in its choices and misclassifing citizens that happen to share this feature as highly likely to be fraudsters.

One the flip side of this relationship stand the rightful benefit receivers and their advocates/supporters. This group does also engage in forms of ”gaming”. Their intentions are rightful and just. Their interest is to receive the highest possible amount of benefits they are entitled to. Under this premise they exert influence onto the system by utilizing methods within their legal rights to receive the highest possible amount of benefits to support their families and living situations. This tactic is deployed due to necessity, since even the highest benefits amount are often barley enough to maintain a minimum living standard. Unfortunately, this form of ”gaming” may sometimes be received as indicators for fraud by the governmental system. We could see this unfold in the toeslagenaffair, were child care organisations helped parents to apply for benefits. Believing they were acting within their legal latitude, they tried to obtain the highest amount of benefits. Unfortunately, the FSV and surrounding systems did not recognize this instance and classified these cases in the same ways, real fraud cases would be classified. From this we learn another important cause for the toeslagenaffair: a lack of support and education during the application for benefits. Measures in this area could possibly make it easier in the future to distinguish the ’intent’ behind a benefit application.

Regardless, these two examples show the clear effect ”gaming” has on governmental systems, and the fact that it can not be underestimated as an environmental conditions, impacting system safety. In feature studies it might prove prudent to investigate the implications of gaming theory (Axelrod &

Hamilton, 1981) for ADM systems utilization in modern social welfare administration. This however falls out of the scope of this research.

Private Third party developers
Often as in the example of the toeslagenaffair, private third party developers are contracted to delivery some of the sub-systems. In the case of the toeslagenaffair, the Belastingidienst contracted third party developers to develop the FSV Dagboak Pit application. While there are many benefits to the involvement of third party developers from the private sector, there are also some negative drawbacks that if not address create potentially hazardous system states. The private sector brings along a number of benefits such as technical know how, project management knowledge, a business orientated mindset etc. However private companies, also implement a different culture into the process. Their workforce is project directed, not necessarily long term system orientated. Third party developers are not as concerned with long-term operations of a system, since this is not a job they will eventually have to do. Their focus is on achieving the goals set in the contractual agreement with the government with as minimal resources as possible in the provided time schedule within the set quality regulation. This results in the nature of the contractual agreement being pivotal to the overall project effectiveness. In practices this step is often rushed leading to a fragile base for the subsequent project to build upon. The nature of the contractual agreements makes it so that a hand-off is created based on predetermined requirements, that are however to technically focused and subject to minimal change. The rigidness of this contract results in inaccurate assumptions being integrated into the system. Additionally projects accompanied by third party developers are often more prone to grow in size and complexity then to shrink. This is due to the third party organisations objective to “increase business”. Therefore they are interested in selling new features to clients, while simultaneously maintain a large deployment size, since this is better then a long-term commitment plan to co-development with a feature plan of 5+ years. Appendix D exemplifies the extensiveness of the use case size on the example of the FSV system. Apart from this mental focus, the contracting third party developers are also likely to not have a vested long-term interest in a stable operability process. They are measured on their ’delivery’ and the design is ’handed-off’ to the client. After the responsibility has passed to the client the resulting issue do not represent a problem but rather an opportunity to the third party developers. After this point the project based resources are not available to the system designers any more and they must be rehired with costly change request agreements. Often this can also evoke the culture of “getting through deployment” or system hand off within teams from third party developers. Errors passed to operations however is a core issue of low quality IT systems (Kim et al., 2016b). The client on the other hand is unlike to committee to costly change request that jeopardize short-term operability. Hence the system is not changed and left to deteriorate over time.

System complexity of government
This topic has been extensively covered in chapter 1. Yet it is necessary to reiterate at this point in time that the overall system complexity that administration has to content with substantial. The relevant factors will not be reiterated, however it is clear that modern government entails a number of highly complex sub-systems that in some cases follow diverging objectives, how in some cases stand in conflict with each other. Finding the right trade-offs, especially in relation to safety is no easy nor trivial task. Building systems in this context we must almost assume failure, at the least we can not assume perfection. Consequently, continuous improvement mechanisms are vital in order to ensure we learn form our errors and continue to thrive towards good governance for all.

## 4.5. Conclusion

The first sub-research question aimed to identify the underlying reasons for the emergence of digital cages, using STPA on the exemplary case of the Dutch child care benefit scandal (Toeslagenaffair). To this purpose first, a stakeholder and an institutional analysis were conducted in order to delineate the system boundaries and explain the environmental conditions surrounding the system. Subsequently, a system safety analysis using STPA was conducted. Through this process, the research identified unacceptable loss scenarios for the example case and moved on to analyze the control diagram and hierarchical control structure of the Belastingdienst system present at the time of the childcare benefit scandal. From these illustrations unsafe control actions plaguing the system were identified. These unsafe control actions build the foundation for the system hazards present in the system. From this, a general inability to identify, control, and recover system states could be identified in the system. Particular emphasis should be given to the system failure to measure systems states through correctly implemented control action and its lack of feedback to system controllers, on theses system state. Additionally, we observed that the issues present in the system are of devers nature both in their respective location of occuarance in the system but also in their type. Some hazards are of technical nature such as a clear diterance in data quality of data registration system in the system, others of organizational nature, such a clear misalignment of mental models between organisational silos that perpetuated miscommunication. Similarly, the analysis also points towards the multitude of environmental conditions that help perpetuate the hazards into harms. Lastly, the hierarchical control structure also illustrates the magnitude of the complexity present in this system and thereby foreshadows the extensive scope needed to tackle safety issues in such a context.

# 5

# Constraint & Requirement elicitation

Within this chapter the second research sub-question is covered: *"What are objectives and requirements for ADM system safety control interventions to curb digital cages in dutch social welfare administration, derived from the example of the toeslagenaffair?"*. Consequently, this chapter provides an elicitation of the intervention requirements derived from the previous STPA analysis and thereby carries out an important sub-step of the STPA analysis, which is to derive system constraints out of the previously derived system hazards. This will be done in section 5.1. Hereinafter, this chapter moves on to discuss general requirements that the interventions should follow, to comply with system safety theory as well as to maintain applicability for other contexts.

## 5.1. STPA: System Constraint Objectives

### 5.1.1. Perception of identified System Hazards

This subsection discusses our perception of the hazards that we have identified in the previous chapter. This is done to firstly acknowledge that while the utilization of scientific methods is applied to prevent personal biases from influencing research results, it can rally be fully eliminated. Secondly, we seek to provide more context that will enable the reader to better comprehend our subsequent choice of system constraint to guide the artefact design.

From the previously delineated unsafe system states and corresponding system hazards we can defer several issues that, in our view, these can be attributed to one of the following mistakes: (1) Making incorrect assumptions about the system during design phase, (2) Not knowing , measuring, or wrongfully interpreting system state, in relation to safety and (3) not being able to effectively adapt/recover system to a safe state after unsafe state has been detected.

In our view these mistakes were made due to unawareness regarding a number of relevant aspects. - unawareness of possibility of digital cage, resulting in safety as a low priority - misconception about system of systems, as static, though it is always changing - unawareness of the curse of flexibility and the corresponding need for 'adaptability' and 'mitigation'

Absolutely! Here's the revised text with the additional information added to the first point:

The first identified issue, of making incorrect assumptions about the system during the design phase, often stem from an idealistic perspective that underestimates the potential for harmful behaviors within the system, such as the formation of digital cages. A significant contributing factor is that these systems are frequently designed in a "vacuum," meaning that assumptions are not verified or tested extensively prior to system deployment. This lack of validation exacerbates the issue as safety considerations become deprioritized. The underlying cause of this issue is generally a lack of awareness of the potential for digital cages to emerge, which results in safety considerations being deprioritized. Designers and policymakers might not have anticipated how rigid and algorithmically inflexible systems could trap users in detrimental scenarios, leading to systemic issues that were overlooked at the foundational design level.

The second issue, not knowing, measuring, or wrongfully interpreting the system state in relation to safety. There is often a lack of adequate tools and processes to effectively monitor or interpret the ongoing state of the system, which leads to an inability to recognize unsafe states as they develop. This issue often arises from a misconception about the ADM systems as parts of a **static** "system of systems." Such a static view fails to account for the dynamic and continuously evolving nature of these systems, leading to insufficient monitoring and adaptive response mechanisms. Without an understanding of the complex interdependencies and the evolving nature of system components, accurately gauging or responding to emerging threats becomes challenging.

The third issue is the inability to effectively adapt or recover the system to a safe state after an unsafe state has been detected. Once problems are identified, the system's rigid structure and predefined protocols hinder effective adaptation or recovery, preventing the resolution of detected issues and the restoration of safe operations. This primarily results from an "unawareness of the curse of flexibility" (N. Leveson, 2011) and the corresponding need for adaptability and mitigation strategies. The system's design does not accommodate the necessary ability to adjust to changing conditions or rectify detected faults, lacking mechanisms for adaptability and mitigation to manage and rectify unsafe conditions effectively.

To address these critical issues, the thesis advocates for the integration of constraints in the design of ADM systems. These constraints should help to ensure the system does not 'migrate to a higher level of risk' overtime (N. Leveson, 2011). The constraints aim to enhance the safety and responsiveness of ADM systems within social welfare, ensuring they are better prepared to handle the complexities and dynamics of real-world environments and prevent future occurrences similar to the Toeslagenaffair.

### 5.1.2. System Constraints

As previously established 'constraints' are a pivotal concept in system safety theory as they represent the mechanism through which emergent properties of a system can be regulated, as impose restrictions on the degree of freedom of the lower level components that evoke the emergent behavior (N. Leveson, 2011). Constraints are imposed onto system components through instances placed higher in systems hierarchy (N. Leveson, 2011). Choosing the right level of abstraction for specific system constraints is vital. Are the constraints placed to low into the system structure, they may not prevent the undesired emergent properties entirely, while constraints that are placed to low do not provide the right specificity. Insufficient specificity can either lead to too many degrees of freedom still available to the sub systems or to an insufficient amount of freedom, also resulting in sub optimal solutions in both system safety and performance terms. This best illustrated by an example. Once could for instance evoke the constraint to "not utilize ADM systems in social welfare administration". While this constraint would solve any safety related issues with automation in social welfare it is not per se a system state to be eliminated or prevented (otherwise, we would not conduct this analysis). For all the issues related to digitization of social welfare administration, there also benefits, and areas were automation makes perfect scenes. Hence the use of ADM is a state that the modern social welfare system must normally be in to accomplish its goals.

Hereinafter the system constraint resulting out of the system hazards and sub-hazards will be presented. The respective constraints do not require a detailed explanation as they can be inferred in a straightforward manner from the respective system hazard. Table 5.1 lists the respective system constraints.

This research is directed towards improving system safety for social welfare administration utilizing ADM systems. Hence enabling safety constraints to effectively restrain the degrees of freedom of lower level system components and thereby make the system safer is the overall objective of this research. This leads to the system constraints reenchanting into design objectives. The constraints are formalized as objectives, because objectives are more flexible and can be negotiated or adapted as the project progresses. They are directional hence provide guidance on the general direction or intent of the design. Objective are often more qualitative than quantitative, though they can be measured in terms of progress towards the goal. The constraints are aspirations, meaning we are thriving to achieve these constraints, yet they represent the ideal outcomes or goals that stakeholders hope to achieve.

| System Constraint ID | | System Constraint | Hazards | Losses |
|---|---|---|---|---|
| **[SC-1]** | | **System must ensure citizens are not misclassified** | **[SH-1]** | **L1** |
| | Organisation | | | |
| | [SC-1.1] | System must maintain effective communication between subsystems | [SH-1.1] | |
| | [SC-1.2] | Signals must be verified prior to processing | [SH-1.2] | |
| | [SC-1.3] | System must maintain infrastructure and processes to determine the current system state | [SH-1.3] | |
| | [SC-1.4] | Mental model alignment must be contiously practiced and documented | [SH-1.4],<br>[SH-1.11] | |
| | [SC-1.5] | Differences in data gathering and registration must be insured not to influence subsequent process decisions | [SH-1.5] | |
| | Data | | | |
| | [SC-1.6] | A sufficant data quality standard for every data registration system must be determined and maintained | [SH-1.6] | |
| | [SC-1.7] | A balanced culture grounded in the principals of good governance must be maintained | [SH-1.7] | |
| | [SC-1.8] | A process for data structure adaptation and improvement must be estabished | [SH-1.8] | |
| | Model | | | |
| | [SC-1.9] | Deployed models must be visable to information handlers and provide explenation for output signals | [SH-1.9] | |
| | [SC-1.10] | Model and process decisions must be traceable | [SH-1.10] | |
| | [SC-1.11] | System must establish causality prior to making decisions on the bases of the output signal generated by the model | [SH-1.11] | |
| | [SC-1.12] | System must ensure fairness and unbias in ML/AI design cycles | [SH-1.12] | |
| **[SC-2]** | | **System must provider citizens with the possible to seek support, taking their respective circumstances into account** | **[SH-2]** | **L1, L2** |
| | [SC-2.1] | A balanced culture grounded in the principals of good governance must be maintained | [SH-2.1] | |
| | [SC-2.2] | If during appeal process, decision can not be reconstructed a process for reassesment must be in place | [SH-2.2] | |
| **[SC-3]** | | **If data is shared with or received from third party organisation, measures must be take to ensure correct usage** | **[SH-3]** | **L1, L2** |
| | [SC-3.1] | If misclassification of externaly forwarded data occures, a recovery process to mitigate impact must be in place | [SH-3.1] | |
| | [SC-3.2] | If insuffianct data quality is detected, data must be ”cleaned” and mitigation actions must be deployed to prevent any losses | [SH-3.2] | |
| **[SC-4]** | | **If system is detected to have failed or throws an error, a recovery correction protocol must be in place that brings system back into a safe state** | **[SH-4]** | **L1, L3, L4** |
| | [SC-4.1] | Maintain modular system and process models allowing for ”function strip down” | [SH-4.1] | |
| | [SC-4.2] | System must maintain effective communication between subsystems | [SH-4.2] | |
| | [SC-4.3] | Implement a process of mediation that ensures effective system utilization across all use cases the system fulfilles | [SH-4.3] | |
| **[SC-5]** | | **Because adaption and migration to a higher risk level are a constant risk, contious development resources for process and product improvement must be inplace** | **[SH-5]** | **L3, L4** |
| | [SC-5.1] | System must maintain clear standards for development support and provide sufficant feedback to development teams | [SH-5.1] | |
| | [SC-5.2] | System must provide sufficant development resources for system adaption and improvement | [SH-5.2] | |
| | [SC-5.3] | Feedback process and change request must be streamlined to ensure longterm effective system operations | [SH-5.3] | |
| | [SC-5.4] | Design chosises in development must be assest based on overall required functionality idependant of individual client interests | [SH-5.4] | |
| **[SC-6]** | | **The System state must be known and attainable at any point in time** | **[SH-6]** | **L1, L2, L3** |
| | [SC-6.1] | System must maintain functionable feedback loops | [SH-6.1] | |
| | [SC-6.2] | System must measure system state at key work stations | [SH-6.2] | |
| | [SC-6.3] | System must insure reports about system state are interpreted corretly | [SH-6.3] | |

Table 5.1: System Constraints that build the Intervention Objectives

## 5.2. Requirements for the Intervention:

To finalize the Requirement elicitation, we turn to the requirements that can be directed towards the nature of the design interventions itself. These aspects will build the scaffold within which each of the Interventions can be built.

### 5.2.1. Hierarchical Levels of Intervention

The importance hierarchy theory to system safety constraints as been outlined in section 5. Since the aim of the safety interventions should be to promote the enforcement of the safety constraints considering hierarchy for the interventions is important. As such hereinafter we will briefly outline the different hierarchical levels that the interventions should address.

- **Institutional and Policy Level:** Rules and regulations that govern the overall operation of social welfare ADM systems. This level sets the broad objectives and legal frameworks.

- **Organizational Level:** Specific organizational practices and procedures that implement institutional policies. This includes protocols, training programs, and oversight mechanisms.

- **Technical/System Level:** Direct interventions into the ADM system and its components. This involves software updates, hardware modifications, or the integration of safety features.

- **Operational Level:** Day-to-day operation procedures, user guidelines, and intervention protocols directly related to the use of the ADM system.

### 5.2.2. Specificity vs Generalizability

Just as constraints need to balance specificity so too do interventions. While interventions should always be made as specific as possible in our case we must consider the utility loss that maybe incurred through providing highly specific interventions. Since the example case we have utilized is a case form the past, inferring highly specified interventions for this case does not yield high utility as the system does not exist in this instance any longer. The goals was always to derive system constraints, from a known case of digital cage formation, that would be generalizable towards other, similar ADM application use cases. This has been mentioned on multiple occasions in this thesis and is the argument for why this research chose to apply STPA not CAST, see 4.4 also 3.3.3.

### 5.2.3. Actionable & Enforceable

The interventions should be actionable, outlining clearly defined the steps, responsibilities, and resources required to implement each intervention. They should also be enforceable, establishing mechanisms for monitoring compliance and effectiveness of interventions, procedures for adjustments.

## 5.3. Conclusion

This chapter seeked to provide an answer to the second research question by, providing design requirements that help inform and evaluate the subsequent design. This has been accomplished by deriving system constraints that can be traced to system hazards and corresponding losses that would be unacceptable to system stakeholders, see 5.1. These constraints help to constrain the system behaviors to a ”safe system state” and thereby represent objectives for the subsequent intervention to achieve, maintain, or recover. Additionally, several requirements regarding the general nature of the design intervention have been derived, these aspects have to be considered in order to maintain the applicability of the intervention for similar cases, since the Toeslagenaffair only represents an exemplary unit of analysis for this process. Finally, the requirement elicitation illustrates the importance of a holistic and exhaustive system safety analysis, issues that have been previously identified and analyzed can be incorporated into system constraints/requirements with limited expense. However, this inadvertently results in an extensive list of requirements that can not be fulfilled by a one-dimensional solution approach. Hence the following chapter introduces a layered approach to the intervention design.

# 6 Control Interventions Design

This chapter addresses the third sub-question: *”How can safety control interventions for ADM systems in Dutch social welfare be structured, formalized, and applied in order to adhere to the identified objectives and requirements?”*. Section 6.1 will initiate the chapter by discussing the problem of enforcing safety constraints in practice and introduce assumptions-based leading indicators as a method by which one can achieve successful enforcement. Section 6.2 will introduce the design of the subsequent intervention by outlining the structure of an assumption-based leading indicators program as the ”main intervention”. Section 6.3 introduces the specific assumptions that build the basis of the leading indicator program. Hereinafter we will move on to the specific interventions that can be used to enforce the leading indicators in section 6.4 to 6.7. The chapter will be concluded through a design demonstration on the example of the Toeslagenaffair in section 6.8.

## 6.1. Issues related to Safety Constraints

The followed STPA analysis has provided a clearer understanding of what existing hazards in social welfare administration that utilizes ADM systems are. The system constraints derived from the analysis provide a good point of departure for maintaining system safety, however the derived system constraints do not come without challenges. It is important we recognize these issues in order to inform our interventions design:

- Complexity: Modern systems, especially those embedded in large socio-technical systems of systems, are incredibly complex. Designing safety constraints that cover all potential failure modes without overly simplifying the system’s functionality can be exceedingly difficult.
- Cost: Implementing robust safety constraints often requires significant investment in both the design and maintenance phases. This includes costs for research, development, testing, and ongoing monitoring. Budget constraints can sometimes lead to compromises that might affect safety.
- Technological Limitations: The available technology may limit the effectiveness of safety constraints. As systems evolve, keeping safety measures up-to-date with the latest technological advancements can be challenging.
- Human Factors: Human error remains one of the most significant risk factors in system safety. Designing constraints that can effectively mitigate risks associated with human error without overly complicating the system or causing operator fatigue is a major challenge.
- Regulatory and Standards Compliance: Adhering to a wide range of safety standards and regulations can be cumbersome. Different regions and sectors may have their own sets of rules, making it challenging for multinational operations or systems that span multiple industries.

- Evolution of Threats: As technology and society evolves, so do the threats to system safety. Systems that deal with external attackers or are utilized in an environment that is subject to ”gaming”, for example, are continually evolving, requiring safety constraints to be regularly updated to counter new vulnerabilities.

- Interoperability Issues: In systems that involve multiple components or are part of larger networks, ensuring that safety constraints are consistently applied and effective across all elements can be difficult. Discrepancies in safety protocols can lead to vulnerabilities.

- Prediction of Unforeseen Scenarios: Despite thorough analysis and testing, it’s often impossible to predict all potential failure modes, especially in complex or innovative systems. Unforeseen scenarios can bypass established safety constraints, leading to failures.

- Balancing Safety and Performance: Managers often views safety constraints and achieving optimal system performance as a trade-off. This hinders the long-term sustainability of safety constraints, as investments into safety perpetuation technology or concepts are lackluster.

Maintaining safety constraints in these systems is therefore a daunting task. Leveson has delinated the previously explained problem with safety and safety constraints as follows:

> ”Too often, system safety is isolated or separated in some way from the system engineering process. The most common result is that safety is treated as an after-the-fact assurance activity. Because safety cannot be assured into a system but must be designed in, safety-related design flaws are often found late, when they cannot be fixed. At that point, the effort then focuses on trying to find arguments that the identified flaws do not need to be fixed. When those arguments cannot be sustained, the efforts to deal with the safety flaws often devolve to making expensive and not very effective solutions, such as redundancy or expecting the operators of the system to detect and fix problems through far-from-ideal procedural solutions”(N. Leveson, 2018)

The results? Often resources are simply not at hand to sufficiently enforce and ensure all safety constraints. This brings us back to an observation made in chapter 1.

That Leveson sums up as follows:

> ”Accidents result from the migration of an organization to a state of increasing risk over time as safeguards and controls are relaxed due to conflicting goals and tradeoffs and reduced perceptions of risk leading to more risky behavior.” (N. Leveson, 2018)

However their are ways we can detect, circumvent and mitigate this causal relation. For this purpose Leveson has introduced the idea of ”Assumption-based leading indicators”. They can utilize within the system safety concept to aid us in this dilemma. The following subsection will delineate this concept further.

### 6.1.1. Assumption-based Lead Indicators for Safety

Leveson outlines leading indicators as follows: they serve the purpose of detecting the potential for an accident before it happens, allowing for preventive actions to be taken. These indicators stem from the assumption that significant accidents are not the result of a singular, random set of immediate events. Rather, they emerge from an organization’s gradual shift towards a riskier system state over time, as the enforcement of safety measures and controls becomes less stringent. This relaxation often occurs due to competing objectives, compromises, and a diminishing awareness of risks, which in turn fosters riskier behaviors (N. Leveson, 2015). Suggesting that the progression towards a major accident unfolds gradually, offers the opportunity to spot this perilous trajectory early and take corrective action. A leading indicator, thus, acts as a signal that intervention is needed (N. Leveson, 2018). Leveson distinguishes between three general types of assumptions that lead to the migration to a higher degree of risk (N. Leveson, 2015):

- 1. The models and assumptions used during initial decision making and design are correct.

- 2. The system will be constructed, operated, and maintained in the manner assumed by the designers.

- 3. The models and assumptions are not violated by changes in the system, such as changes in procedures, or by changes in the environment

For organizations, leading indicators thus offer primitive signals that elements of their products, services, or behaviors may be beginning to deviate from intended system states (N. Leveson, 2018). They therefore present a great opportunity to simply the process of maintaining system constraints, and preemptively curb the emergence of digital cages / accidents. The methods hinges on the idea that understanding the underlying assumptions pivotal to the safety design of a specific organization, product or operation will greatly aid in identifying reliable leading indicators for risk migration (N. Leveson, 2015).

> ”No engineering process is perfect nor is human behavior. In addition, every system and its environment are subject to change over time. The starting point in seeking more effective leading indicators is to consider the general causes of accident”. (N. Leveson, 2018)

### 6.1.2. Assumptions underlying System Constraints as bases for Leading Indicators

The tie between safety constraints and assumption-based leading indicators lies in their mutual goal of preventing system failures and ensuring safety. Safety constraints provide a foundation for identifying critical assumptions and developing leading indicators. By understanding the boundaries that define safe operation, system designers and operators can pinpoint which assumptions are critical for maintaining these constraints. As illustrated above assumption-based leading indicators offer a proactive approach to safety management. By monitoring these indicators, organizations can identify when a system is approaching its safety constraints and take corrective action before a constraint is violated, thereby preventing accidents or system failures. As leading indicators provide early warnings of potential safety issues, they can inform the operators when their is an increased risk that safety constraints may be violated. The also point towards the need to revisit and possibly revise safety constraints to better reflect the current understanding of the system and its environment, since once an assumption has ”failed” the integrity of the underling safety design of a system is threatened(N. Leveson, 2015). Together, safety constraints and assumption-based leading indicators support a culture of continuous improvement in safety management. By continuously monitoring leading indicators and reassessing safety constraints, organizations can adapt to changes in the system’s operational environment and emerging risks, enhancing the overall safety of the system. Subsequently, eliminating the underlying cause for the emergence of digital cages in deployment and operation, provided the leading indicators and assumptions are implemented correctly.

Therefore, assumption-based leading indicators will build the bases for the design of the subsequent interventions. It hence can be viewed as an overarching process encapsulate and deliver each intervention as part of something bigger. Introducing the leading indicators as a method to improve ADM system safety in social welfare, also allows feature researchers and practitioners to adopt other specific forms of interventions to incorporate into this structure and utilize for the specific cases they are analysing.

## 6.2. Design of Intervention

The design our intervention by analysing our system constraints. Each of the constraints is based on underlying assumptions that the system is able to execute specific functions in order to maintain the constraint. We start with identifying the most pivotal assumptions. These will be used to generate safety-based lead indicators that help system operators to verify assumptions, detect and interpret unsafe system states and aid in a safe resolution of emergent hazardous system states (N. Leveson, 2018). Hereinafter we identify ways by which we can enforce the leading indicators onto the system. There are three ways in which lead indicators can be enforced (N. Leveson, 2018):

- **Shaping actions** to prevent violation of the assumptions
- **Hedging actions** to prepare for failure of an assumption
- **Assumption checking during operations**

  Sign posts to trigger specific checks

Checking during system operation (periodic or continual)

Performance audits

Surveys

Automatically collected data

For each of these three modes of enforcement we will propose specific interventions. To finalize we will look to group the interventions, to look for synergies in assumptions. While in an ideal case one should seek to verify all assumption that are made, in reality restraining the amount of lead indicators keeps the process comprehensible. Moreover inferences between assumptions maybe deducted, meaning we can reduce the amount if assumptions that need to be monitored by deducting that if assumption 'A' is measured to have failed, the likelihood that assumption 'B' & 'C' are also violated is high. This step is also revered to as creating a "leading indicator monitoring program" (N. Leveson, 2018). Since detection and prevention alone, are not sufficient, there must also be a "management process" in-place once leading indicators are triggered (N. Leveson, 2018). This will be done through providing a "Hedging process" described in the intervention section. The "leading indicator monitoring program" will be formulated during the "design demonstration" phase (Johannesson & Perjons, 2014), later in this chapter. Because creating a monitoring program needs to be directed towards a specific system, in our case the previously analyzed Belastingdienst/Toeslagen system.

## 6.3. Assumption-based Leading Indicators

Within this section we will outline the specific assumption-based leading indicators that can be derived from the system safety constraints from the previous chapter. This means the assumption are grounded within the hazards identified in the toeslagenaffair, however there applicability research beyond this specific case. The prior sections have outlined the *Why?* the leading indicators are needed. This section will outline: *What specific assumptions can be derived?*, while the following sections will discussed the question *How can the assumptions be enforced?*.

| ID | Assumption-based Lead Indicator | Constraints |
|---|---|---|
| 1 | Assumes it is feasible System is able to always classify individuals correctly | [SC-1] |
| 2 | Assums that misclassification in system is detectable | [SC-1], [SC-3], [SC-4], [SC-5], [SC-6] |
| 3 | Assumes a 'normal' system workload & rate | [SC-1], [SC-2], [SC-6] |
| 4 | Assums effective communication between sub-systems | [SC_1-6] |
| 5 | Assums Mental model alignment throughout process and system | [SC_1-6] |
| 6 | Data quality in registration and models is assumet to be of high quality | [SC-1], [SC-3] |
| 7 | Assumes that incooperated feedback functions are executed | [SC-2], [SC-3], [SC-4], [SC-5, [SC-6] |
| 8 | Assumes that system state is measurable | [SC-1], [SC-4], [SC-1], [SC-6] |
| 9 | Assumes that system data is interpretable | [SC-1], [SC-4], [SC-1], [SC-6] |
| 10 | Assumes that sufficant system state measurments are made | [SC-1], [SC-3], [SC-4], [SC-5], [SC-6] |
| 11 | Assumes that there is a structural organisation facilitating this process of control | [SC_1-6] |
| 12 | Assumes that thrid party organisation provide accurate data | [SC-3] |
| 13 | Assumes that thrid party organisation follow provided data ontology | [SC-3] |
| 14 | Assumes colleted model data to be unbiase | [SC-1] |
| 15 | Assumes a safety orientated culture grounded in principals of Good Governance | [SC_1-6] |
| 16 | Assumes enough development resources to adapt system | [SC-4], [SC-5] |
| 17 | Assumes that the appeal process is functional and mitigates arising harms | [SC-4], [SC-6] |

Table 6.1: Assumption-based Leading Indicators, derived from the System Constraints

These assumptions underpin the expectations about system behavior and the parameters within which the system is intended to function safely and effectively. For instance, the assumption that "it is feasible, System (not model, system) is able to always classify individuals correctly" challenges the system designers to ensure the ADM systems are not only accurate but also have mechanisms in place to detect and correct errors, thereby preventing misclassifications that could trap individuals in digital cages. Similarly, the assumption about "misclassification in system is detectable" highlights the need for robust detection mechanisms that can identify errors in real-time, an essential feature for maintaining system integrity and trust.

Further, assumptions such as "effective communication between sub-systems" and "mental model alignment throughout process and system" emphasize the importance of coherence and transparency across the system's architecture. These assumptions ensure that all parts of the ADM system are harmoniously integrated and that stakeholders have a uniform understanding of how the system operates, which is crucial for both system reliability and user trust.

Data-related assumptions like "data quality in registration and models is assumed to be high" and "assumes collected model data to be unbiased" are particularly significant. They underline the necessity for high-quality, unbiased data as the foundation for decision-making processes, directly impacting the system's fairness and efficacy. Ensuring data integrity and addressing potential biases are fundamental to preventing the emergence of digital cages, where individuals might be unfairly treated due to flawed data inputs.

The assumption that "the appeal process is functional and mitigates arising harms" integrates a crucial safety net within the system, providing a mechanism for redress and correction of system errors, thereby enhancing the system's fairness and accountability.

Each of these assumptions, represented through specific leading indicators, serves as a proactive measure to predict and mitigate potential system failures. They ensure the system operates within its safety constraints and maintains its intended functional integrity over time. The next sections will explore practical methods to enforce these assumptions effectively, ensuring continuous safety and reliability of the ADM systems.

## 6.4. Interventions: subsequent structure

In an effort to answer the question *How can the assumptions be enforced?* the subsequent sections will discuss interventions, grouped into their respective 'mode of enforcement'. We will firstly discuss the interventions by which assumptions can be checked within operations, then move towards 'shaping actions' and finally discuss 'hedging actions'. The Interventions are introduced in similar fashion. The chapters seek to provide answers to the questions: "Why is the intervention needed?": through connecting it to previously discussed issues, "What is the intervention?": outlining its conceptual scaffold, and "How can the interventions solve the issue?" explaining how the intervention may alleviate the respective issue. To allow for applicability to other scenarios, contextualization of the interventions to the Toeslagenaffair, will be kept minimal. Only using the Toeslagenaffair to illustrate specific points. The interventions will be contextualized more in-depth in the section 6.7 "Design Demonstration". Within that section answers to the "Where?", "When?" and "Who?" questions for the case of the Toeslagenaffair will be provided.

## 6.5. Checking assumptions in Operations

Hereinafter, we will outline the respective interventions that can be used to check assumptions during operations. The are highly critical interventions, ultimately one can only account to things one can measure and interpret. Hence checking that assumptions are valid in operations represent the core activity of an "assumption-based leading indicator" program. We must ensure that the system is in a safe operational state, by making sure its underlying design assumptions retain their validity and integrity. The subsequent interventions, may also be used in a different capacity and my for example also be utilized in as a "shaping action". If such a duality applies, this will be mentioned. For purposes of consistence and readability the respective interventions will not be reintroduced in the other corresponding section "[...] action" section. For the assumption checking during operations, we can summarize that its all about identifying the system state. In the world of bureaucratic decision making there are often no physical boundaries, in comparison to for example an aircraft were an artefact is controlled in a physical setting. This makes it very difficult for controllers to derive an accurate system state, because (1) the system controls a 'cognitive process' (Jakubiec, 2022) (2) the system boundaries are ambiguous. This aspect emphasizes the importance of deriving good leading indicators, and their importance. Subsequently, we will outlines means by which these leading indicators may be verified within the complex system environment of social welfare administration.

### 6.5.1. Telemetry

Why?

In the handbook the authors describe how a general rule of thump for software issues in deployment was to reboot the server on which an error occurred. This was because their was no information available about what could have caused the error, it could have been anything from a failure in application, environment or externally induced (Kim et al., 2016b). This is similar to the situation that occurs in current social administration systems after a system error, misclassifcation has been detected, because their is not further information available controllers are not able to understand what sub-system or process might have caused the error. This can be refereed to not being able to detect, measure and interpret the current system state.

What?
An intervention that can be derived from the DevOps Handbook is telemetry (Kim et al., 2016b).Telemetry as defined in the book is ”an automated communication process by which measurements and other data are collected at remote points and are subsequently transmitted to receiving equipment for monitoring” (Kim et al., 2016b). This process bears in and of itself a lot of resembles with the leading indicators, since telemetry is also trying to identity indicators in the data that may point to the underlying cause of en error. Leading indicators are however a broader concept dealing with the issue of system safety. Nonetheless, telemetry can be viewed as a ’checking action’ or ”shaping action” ensuring the integrity of underlying assumptions.

How?
Telematry allows system controllers to collect ample information about the sub-systems state, build inferences regarding the interconnections and understand metadata that the systems may provide us with. Telematry can be used as a method to check assumptions during operations, because we can use it to report data directly relevant to lead indicators. It can be used as a ”shaping action” by using the information inferred from telemetry as bases to improve our system safety design through for example utilizing it as the bases for a continuous improvement process. As illustrated in figure 6.1, the information collected should not only entail information about operational sub-systems, it should also include information about the status of applications running on these systems, as well as signals that are related to business logic. Through telemetry the system controllers should measure core metricises relevant to the leading indicators outlined in section 6.3. Among them for example the workload experienced by the different sub-systems. The amount of decision that are made in a specific time frame, incoming and outbound signals. This should enable the controllers to visualize ”events” in the system from a more holistic perspective build inferences from the received information that helps to determine the current system state.

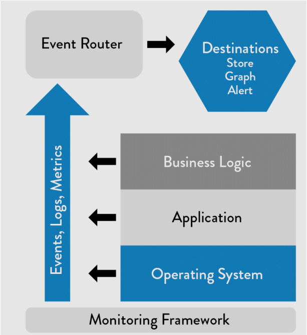


Figure 6.1: Telemetry Monitoring framework from DevOps Handbook(Kim et al., 2016a)

### 6.5.2. DevSafOps Teams

Why
Through the STPA analysis of the toeslagenaffair the importance of mental model alignment is emphasized. It is clear that the misalignment regarding the state of the process and the meaning of specific information’s/status played a big role enabling the ’ripple effects’ that ultimately lead to the digital cage. Specifically, for toeslagen, the misalignment regarding the ”1x1” boxes exemplifies the immense impact, minor misalignment’s can cause if pared with environmental worst case conditions. Figure 6.2 illustrates the different mental models existent and their respective relations within a system (R. Dobbe et al., 2021). For the case of the toeslagenaffair, the mental model of different operators within the

system was not aligned with each other. One can imagine the that such a misalignment inadvertently will also lead to confusion and misalignment with the other actors depicted in the illustration. Consequently, mental model alignment must be continuous in a complex system, since otherwise the actors will "drift".

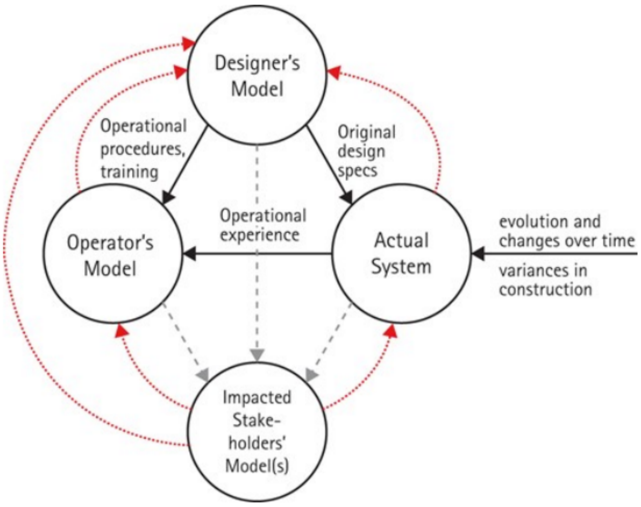

Designer's Model
Operational procedures, training
Original design specs
Operator's Model
Operational experience
Actual System
evolution and changes over time
variances in construction
Impacted Stake-holders' Model(s)

Figure 6.2: Mental Model alignment(R. Dobbe et al., 2021)

What?
One intervention that has the potential to improve mental model alignment is "DevSafOps". An evolution of the "DevSecOps" terms, reaffirming the need to shift towards "safety" not "security" in the context of ADMs in social welfare administration. Safety as a concept, entail a broader idea and does not merely focus on threats to the system but also "threats" the system itself my create. Bringing Development, Safety and Operations into once interdisciplinary team that views a process from a holistic perspective makes mental model alignment possible. Figure 6.2 illustrates where "DevSafOps" may conceptually be placed.

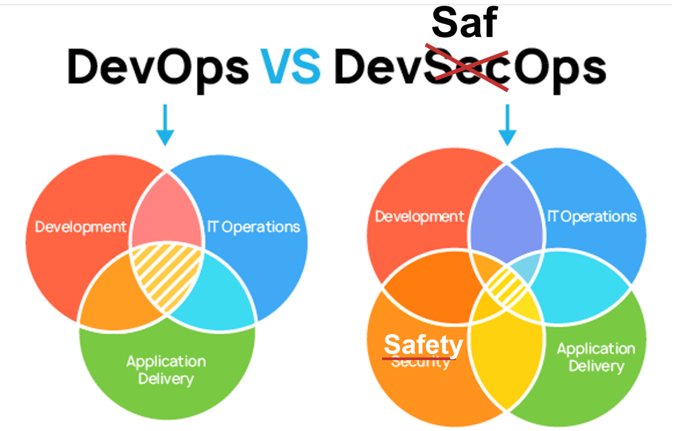


Figure 6.3: DevSafOps as a new organisational structure(Kim et al., 2016a)

How?
Bringing Development, Safety and Operations into one interdisciplinary team that views a process from a holistic perspective makes mental model alignment possible. These new structures should use the organisational principals outlined in the DevOps Handbook as guidelines (Kim et al., 2016b). This

means the team should be distributed throughout the different sub-process of the system and keep constant communication with each other. Ultimately this can contribute to fulfill the following functions:

- Cross-Functional Collaboration: DevSafOps encourages close collaboration among development, safety, and operations teams. This integrated team structure can also include domain experts from the system's functional areas, ensuring all perspectives are considered in the development and operation of the systems. This helps in aligning mental models across different stakeholders such as developers, operators, domain experts, management and users organisations on how the system should function and its intended outcomes.
- Iterative Development and Feedback: DevSafOps promotes iterative development, which allows for frequent testing and feedback incorporation. By utilizing this approach, assumptions can be continuously tested and refined through stages like development, deployment, and post-deployment. This iterative loop helps in aligning the mental models of the software with real-world administrative needs and expectations.
- Transparency and Documentation: Clear and accessible documentation plays a crucial role in maintaining an aligned understanding across the team. Documenting assumptions, decision-making criteria, data sources, and algorithms used in ADM systems helps ensure that every team member understands the system's mechanics, thereby fostering a common mental model.
- Testing and Validation: Security and compliance testing, along with performance evaluations, can uncover discrepancies and implicit assumptions in the system. Testing scenarios can be designed to simulate real-world conditions under which the system will operate, providing insights into how well the system's assumptions hold up against real world use.
- Training: Providing training sessions and workshops for all organisations involved with the systems processes helps to establish a shared understanding and alignment of mental models. These educational initiatives should focus on the safety constraints and assumptions underlying the system, the nature of deployed technology (e.g. Causality in ML) and the significance of data quality and integrity.
- Continuous Monitoring and Adaptation: Once the System is deployed, continuous monitoring can help track its performance against expected outcomes and assumptions. DevSafOps could play a pivotal role in interpreting this real-time data and provide feedback that can be used to refine and align the mental models continuously.

Focused on these objectives the teams could be centered around the ”assumption-based leading indicator” program. Tasked with verifying their validity in the operational process.

### 6.5.3. Audit

Why?
The issue of adaptation has been discussed within chapter 1. Adaptation is once source of digital cage emergence and often manifests itself in form of ”deviations” from an ideal state. Among them for example a shift in underlying culture, which as described in chapter 4, is a hazardous system state that overtime significantly increase the migration towards a riskier system state. Other deviation could be related to moving away from process standards or established quality measurements.

What?
Audits are a great form by which leading indicators can be checked during operations. They provide opportunity to detect misalignment between mental models, as well as audit process standards and other forms of qualitative and quantitative metrics.

How?
There are numerous different audit that might be utilized, heavily dependant on the specific leading indicator that needs to be audited, as well as the organisation and system the audit is conducted in. As such it is difficult to specify this intervention further. However we can point towards established audit practices in public administration, as a guideline for practitioners as to how such audits may be

conducted. These guideline have the added benefits that they are already contextualized towards the specific organisation that utilizes them and that they are connected to an enforcement organisation and process. Ultimately, Audits can contribute to achieve the following objectives, that are either shaping or checking actions:

- Independent Evaluation: Audit teams provide an independent check on the system, verifying that the systems perform as intended and adhere to specified requirements and regulations. This independent assessment helps in identifying discrepancies between the expected and actual behavior of the system.
- Assumption Checking: One of the key roles of audit teams is to scrutinize the assumptions underpinning the system. This includes evaluating the assumptions about data quality, model accuracy, and the suitability of algorithms for the intended use cases. By challenging these assumptions, audit teams help ensure that the system's operations are based on valid and justifiable grounds.
- Risk Management: Audits are adept at identifying potential risks, including those related to bias, discrimination, and privacy violations. Their insights can shape actions to mitigate these risks before they manifest into harms.
- Documentation and Reporting: Comprehensive documentation and detailed reporting are essential for transparency and accountability in the systems. Audits ensure that all processes, decisions, and changes are well-documented, providing a clear trail that can be followed to understand how decisions were made.
- Feedback: Audits provide feedback mechanisms to inform management about the findings and potential for improvement. This feedback is crucial for ensure that the systems adapts in response to changes in the operational environment.
- Compliance and Culture: Ensuring compliance with process and cultural standards is a cornerstone of audit activities. By assessing ADM systems against these standards, they help to ensure that the systems operates in a safe state.
- Training: Audit teams can also contribute to training programs for case workers of respective sub-systems. By educating them on the findings, risks, and operational guidelines, audits can foster a deeper understanding and better mental model alignment.

### 6.5.4. Blue Team vs. Red Team

Why?
Often in the public sector we can observe that the administrative organisation are divided into different divisions that act independently from one another and flow different objectives. The Belastingdienst with its different mandates provided a prim example for this. However, so as in the Toezicht process, often these organisational silos still have to cooperate to fulfill higher organisational objectives. The resulting organisational intersections represent a ”weakness” in the overall system safety, as is illustrated by the amount of unsafe control actions situated within the organisational interconnection in the case Belastingdienst. These weaknesses are difficult to detect yet represent a high risk both in terms of safety and inner-organisational conflict. Therefore they must be controlled meticulously and constant checking regarding their integrity is needed.

What?
The red team vs blue team tactics can be utilized both as a checking as well as a shaping action within the system. Red teams are often utilized in information security to identify weak points that can be exploited. To subsequently find means with which to close these ”holes” and improve system security. Similarly, we can utilize red teams to identify safety related issues.

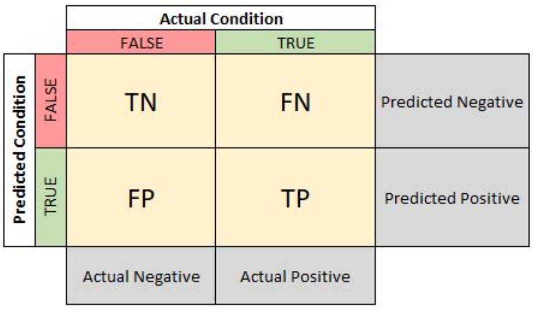
How?
In this context red teams may be used to identify potential weak points for misclassification, or test detection capabilities of the system through implementing wrongful decisions. As such they are great in creating a continuous improvement process. Red teams could best be combined with other interventions such as the DevSafOps team, to create a holistic approach towards safety throughout the systems life-cycle.

### 6.5.5. Dummy Data

Why?
The fact that digital cages are emergent and causality is difficult to trace through the different process steps in the system also create ambiguity in judging whether decisions that have been taken are right or wrong. Is is problematic because it complicates the assumptions verification process immensely. The ambiguity of outcomes makes it difficult to judge the whether the system is in a safety state or not.

What?
Another promising detection intervention during assumption checking in operations is dummy data. With dummy data the system can be check on how it may react to ’boundary cases’. This information is vital in understanding the pitfalls in the system and how the interaction patterns between subsystems may impact how information is processed and interpreted. Dummy data also removes ambiguity out of the assessment, because the expected outcome can be predetermined through a panel of experts.

How?
With dummy data or dummy cases system controllers can feed different sub-processes of the system to determine their respective performance. Figure 6.4 illustrates the confusion matrix that could be utilized to evaluate the performance of the respective sub-process (Hayes, 2015). Controllers can use the confusion matrix to asses the processes and sub-systems for their accuracy, recall (sensitively or specificity) etc. These are important insight that can be used to verify assumptions as well as provide input for shaping action. If for example specificity values are low system controllers could instruct the process to apply stricter criteria, for determining a ”positive” case (Hayes, 2015). Such dummy checks can be us full for both machine as well as human decision making processes within the system. Important to stress here, that it is not a means of identifying errors to assign blame, but to improve long term system safety. Additionally, the length over which these tests are applied can be varied. I might make scenes to test several process steps at once to get a better understanding of how they interface, and to test control and feedback mechanism within these process steps. Consequently, dummy data can be used to both check and shape the systems safety. Coordination of this process should be placed with an interdisciplinary team such as the previously outlined ”DevSafOps” teams.



| | | Actual Condition | | |
|---|---|---|---|---|
| | | FALSE | TRUE | |
| Predicted Condition | FALSE | TN | FN | Predicted Negative |
| | TRUE | FP | TP | Predicted Positive |
| | | Actual Negative | Actual Positive | |

Figure 6.4: confusion matrix (Hayes, 2015)

## 6.6. Shaping actions

Shaping actions are a necessary addition to the leading indicator program, because the system is under constant adaptation. This adaptation may in some cases prove assumption that piously held true, to be invalid. Therefore system operator must deploy shaping actions that help keep the assumptions underlying the system valid or adapt the system to allow for the underlying assumption to change. Shaping action can be highly divers, because they are very system specific. Many of the previously outlined checking actions can also be utilized as shaping actions. Among them for example telemetry, audits, DevSafOps teams and Red Teams.

### 6.6.1. Utilizing System Safety Methodology (STPA etc.)

Why?
The utility of system safety methodology to improve the safety of social welfare administration systems as be exemplified throughout this research. Therefore it comes to now surprise that these methods

should also be used actions to actively shape the systems during their development and operations phases. The need for these methodologies has also been outlined. They provide a holistic system thinking-based approach that enables system controllers to identify, unsafe control action and the corresponding system hazards alive within the system.

What?
The previously deducted STPA analysis only represent of a subset of the possibility that STAMP and STPA can provide for improving system safety. As Leveson points out STPA can be utilized during the entire standard system engineering life cycle, beginning at the initial concept development phase (N. Leveson, 2011). Hence system safety methodology should be utilized in its various forms throughout the systems life-cycle.

How?
STPA can be used to establish initial safety requirements during the concept phase and further refine these during the system requirements development stage. These requirements and constraints subsequently guide the creation of the system architecture and the detailed system design and development. The outcomes of STPA analyses can evolve alongside the design and development phases, offering valuable insights for decision-making (N. Leveson, 2018). Its utility extends into the deployment and operations stages, delivering critical data for operational use and product improvement. Within a model-based engineering framework, STPA operates on a system model that undergoes refinement with design decisions. It enhances traceability throughout the development process, allowing for easy adjustments of decisions and designs with minimal need to redo previous analyses (N. Leveson, 2018). STPA's application is not limited to safety; it can be applied to any emergent property of system engineering and product life-cycle (N. Leveson, 2018). For a clearer delineation of STPA methodology refer back to chapter 3.

### 6.6.2. Batching

Why?
As can be inferred from the documentation available for the ”toeslagenaffair”, more specifically Capgemini's documentation on FSV, the FSV system entailed a large number of system features. These features increase both technical as well a process complexity. Making the deployment of such a system highly complex, extensive in time and resources. We presume that similar developments can be observed ovedr a wide varity of software projects within public administration, for reasons that have been outlined in section 4.4.

What?
From DevOps we learned the need for batching (Kim et al., 2016b). DevOps outlines that keeping deployment as small as possible is the key to success, as it enables a smother process flow, better debugging and easier critical resource management (senior system engineers and developers). Hence in order improve deployment feature complexity of newly developed systems should intentionally kept small. This also mean the underlying processes should be stripped down to their critical functions. Critical functions in this context refer to basic processes that need to be executed to reach some level of system performance. This level has to be detained by the system owner, and is highly dependant on each specific context.

How?
In the example of the FSV system this could for instance entail that the ”1x1” box is added and deployed at a later stage into the system, after the structural bases for the system and the process it facilitates are well established in the organisation function. Deploying in smaller batches make the process of deployment easier reduces organisational confusion and improves the likelihood that possible issues are caught. We acknowledge that in the operational reality of public governance this is not an easy feed to accomplish, because the underlying policy that often triggers the development of new systems, can not always account for making such design/deployment trade-offs. Policy packages are often very large, entailing a number of interdependencies that inadvertently resulting in 'heavy' and opaque system deployments. To address this issue policy should entail instructions on trade-offs for a preliminary

implementation phases of a new policy, to allow the executing organisation to make necessary implementation adjustments without jeopardizing the ”core” nature of the policy. Batching is a shaping action but can also be used as an assumption based leading indicator. In this scenario one would lean on the assumption larger system deployments are more likely to create system hazards, due to their opaque nature.

## 6.7. Hedging Actions

Hedging actions are a necessary component to any safe operation of a system (N. Leveson, 2018). They deal with the potential consequences that are evoked through the failure of assumption. Hedging actions hence are directed towards limiting and mitigating potential harms, as well as returning the system back into a safe state (N. Leveson, 2015). In regards to automated decision making systems in social welfare administration, the overall objective for hedging actions is therefor to prevent the emergence of a digital cage as best as possible, and subsequently guide the administrative bodies in returning to a safety state of operations. It is worth noting that the Hedging actions can only be successful in combination with detection and interpretation interventions discussed previously. Since only when the violation of an assumption is detected, can hedging actions be deployed to their maximum effect.

### 6.7.1. Hedging process

Why?
A problem discussed extensively in section 4.3 is the issue of a missing mitigation process that would have enabled system controllers of the ’Toezicht’ process to return the system into a safe state of operation after an error had been detected. The absence of this process ultimately enabled harms to persists in the system. This is crucial because it enabled the ”digital cage” to emerge fully. Several sources have pointed out that concerns regarding the safety of the Toezicht system had been raised on several occasions prior to the breaking of the scandal in the national news. The instances can be understood as moments, were individuals of groups started questioning the integrity of the assumptions underlying the design of the toeslagen system. However, due to the fact that there was no leading indicator program implemented and now hedging actions deployable to alleviate the raised issues, the typical organisational/human reflex of ”defensive avoidance” (N. Leveson, 2018) was exhibited by system controllers (PwC, 2021). This is exemplified, by manner through which the scandal subsequently unfolded: it took immense public pressure before officials started caving and admitting to the fact that the system was flawed, see 4.1. With clearly defined hedging actions in place such behaviour can be avoided, since clear rules for conduct during system failure are established. Hedging actions, also provide the added benefit of ’protecting’ the decision makers in place, since thru them responsibility is shifted onto a process rather then individuals. Ultimately, one can never assume the design to be flawless and hence correction mechanisms must be in place (N. Leveson, 2011). Procedures and process help guide this process in the safest way possible, if pilots for example had to figure out what to prioritize after an engine failure in flight, without established processes, catastrophe is surly unavoidable (N. Leveson, 2018). Similarly, this also applies to system failure in social welfare. Clear ideas of the priorities, mitigation actions etc. must be established through a predetermined process.

What?
A clearly defined ”hedging process” for the use of ADM systems in social welfare must established. This process should enable the controllers, to mitigate potential harms after an assumption failure has been detected. It has to be structured in a way were the system is returned into a safe state as fast as possible, the issue is resolved quickly and safety of the system is subsequently strengthened lastingly.

How?
For an active hedging approach system controllers may follow the following process outlined in figure 6.5.

Subsequently, the respective sub-hedging actions related to the process will be outlined in the already established fashion.

### 6.7.2. Increased Bottleneck / Critical Process Supervision

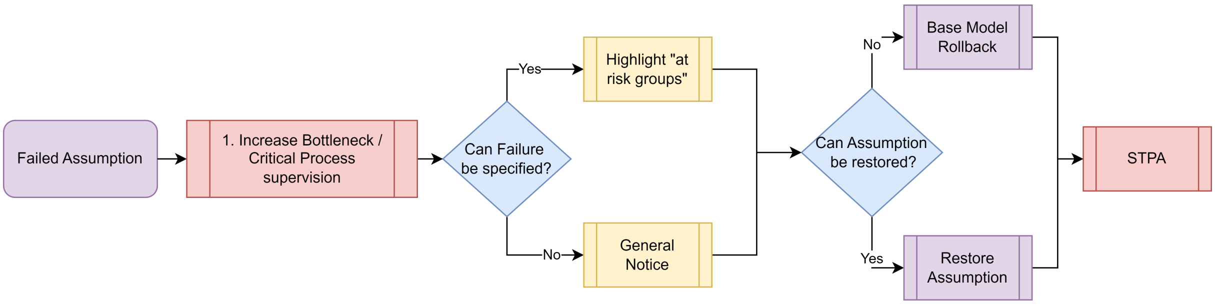

Failed Assumption
1. Increase Bottleneck / Critical Process supervision
Can Failure be specified?
Yes
No
Highlight "at risk groups"
General Notice
Can Assumption be restored?
No
Yes
Base Model Rollback
Restore Assumption
STPA

Figure 6.5: Hedging Action process

Why?
After an assumption has failed the most important task is to avert any negative harms to further converge onto the citizen. This objective is not easy to achieve due to the interconnections that are existent within governance (Peeters & Widlak, 2023). The issue is that often causality is hard to establish in these cases, meaning even if the system controller are aware of the fact that errors and misclassifications have occurred, they are not able to identify who was affected by this system failure. The risk of passing mistakes to downstream work centers is high. As Kim et al. point out in the DevOps Handbook, this inevitably results in quality issues later in the systems workflow (Kim et al., 2016b). Downstream work centers such as the LIC in the example of the Toeslagenaffair, often are simply not able to fully understand the decisions made by previous workstations in the process. They lack the information, time and resources to do so (Kim et al., 2016b). Hence it is vital that once issues have been raised, special care is directed towards making sure that downstream work centers receive accurate, high quality work & information.

What?
From the previously outlined arguments we can delineate that critical processes within the system have to be increasingly supervised. Critical process entail bottleneck processes, meaning processes that any 'case' or process instance has to pass through to reach downstream system processes/workstations, as well as interaction processes with third parties, other parts of the organisation or citizens. The ladder represent the ”last line of defence”, since the interaction with citizens is the process that ultimately converts prior wrongfully decisions into actual consequences and harms.

How?
From this we can deduce that once it is detected that an assumption underlying the system design is broken, these interactions must supervised heavily, or dependant on the incursion even stopped. In practice this means that resources from other organisational functions have to be redirected towards the critical processes and coordination between this processes has to be increased. In the example of the toeslagenaffair, some senior case workers from the Toezichtteam, should be reassigned to the LIC to supervise the ”Invorderings process” (reclaim process). Naturally, some of the organisational decion making directives could be loosened as well in order to allow for a more ”forgiving” nature, which may mitigate arising harms.

### 6.7.3. Highlight ”at risk” Groups

Why?
As stated in the previous subsection, the earbashing causality within the systems failure is difficult. However it remains a goal of system safety, if controllers are able to establish causality the chance of mitigating harms is improved dramatically. The process of establishing causality enables controllers to determine the extent to which the system is compromised, this knowledge can be used to inform feature hedging actions and respond in a ”thread-level aware” fashion.

After an increased bottleneck supervision is triggered the hedging process should move to determine the extent to which the systems design is compromised. Have the underlying assumptions failed completely, or are specific areas 'cases' effected due to a temporary break in assumptions? For example, was there a specific model version that malfunctioned, leading cases were classified by this particular version to be compromised, or are there other discriminative features by which the negative impact can clearly be attributed only to a specific subset of cases?

How
If the system controllers can deduce that only a specific part of the system has failed the respective affected subsets can be highlighted throughout the process as ”at risk” warranting a closer revision. If such inferences can not be made and it must be assumed that the any case instance maybe compromised, a general ”system notice” to all system actors should be given out. In order to detect and answer these questions the previously illustrated interventions under ”operations checking” can be utilized. Crutaly for this process is however the implementation of a process wide logging function, meaning the system must know which sub-models contributed to which decision and at what time the decision was made. Only if this information is know and available the affected subsets be highlighted throughout the systems operational process.

### 6.7.4. Rollback to Base Model

What?
Moving on the controllers have to analyse the severity of the assumption failure. Is it possible with to restore the underlying assumption through minor changes in the process structure, or can the assumption not be restored? In the latter case a system rollback has to be executed. Meaning the system should be rolled back to a simpler system structure / process model that makes it easier to manage the complexity, omitting non critical system features and steps.

How?
In relation to the execution of policy this would mean minor objectives of administration are temporary omitted to avoid breaking higher principals of governance, such as the mandate an underlying principals and laws of good governance. In the example of the toeslagenaffair for instance this could have been achieved by making features such as the ”1x1” box toggable. This would have lead in a decrease of complexity. Granted such intervention would have significantly lowered the systems ability to survey potential fraudsters, yet it also significantly lowered the potential for harm created through the system. Other such omitable processes could have been data received by third party organisations, complex risk models deployed by DF&A or data sharing with thrid party organisations. However system rollbacks should not solely understood as shutting of technical sub-systems and moving the process back to a more ’primitive’ means of operation. It also implies that the process humans within the system followed are stripped down to their ’core functionalities’, otherwise a system rollback would only result in a larger workload for human operators within the system, which is precisely beside the point. A system rollback should see the system focus on its core functionality, clearly prioritizing the most impactful work it is faced with. In the case of the Toeslagenaffair this could therefore also have entailed, that teams nolonger investigate signals that are not deemed as ”high priority”. Rollbacks consequently therefore always entail a trade-off between system ’capabilities’ and ’safety’. However capabilities that can not be executed safely do not make a system more capable.

Why?
Rollbacks allow organisations to ’take a breather’ asses the system state and then trigger a systematic system and process review to identify and resolve safety concerns. Such a rollback should not be triggered lightly, and a list of specified criteria should be work out to determine whether a system rollback may be necessary. As exemplified through the cut backs mentioned for the toeslagenaffair, rollbacks also can entail a significant loss in system performance. Nonetheless, if the underlying assumptions for a system are violated and a risk for digital cages hence is high, such an intervention mitigates the potential for harm.

### 6.7.5. System / Process Review with STPA

Why?
Just as exemplified through its use in this thesis, STPA can be utilized to triggered a system review process. Identifying possible unsafe control actions, finding system hazards, deriving new system constraints and control actions that help to improve the safety of the system. STPA should therefore always be part of the hedging process, to learn from past mistakes and improve the safety of the

system for the future. STPA can also be used to improve the design of the assumption-based leading indicators, to increase the prediction performance and help overall system safety.

What?
After the system been steered back into a safe state a full system review has to be initiated. This is necessary because an assumption failure acts as an indicator for potential problem, however does not fully identify all the issue that might be present within the system. It point however to the fact that the current safety control strucutre is inadequate and has to be adjusted. Therefore a full safety review process has to be triggered that should, orientate itself on the previously outlined methodology of STPA.

How?
The value of an STPA analysis has been exemplified throughout this research, particularly in chapter 4.3. Consequently, we will not delineate how STPA can help to make system safer here. One can add however that for the system review process STPA's most potent capability is its ability to detect unsafe control actions.

## 6.8. Design Demonstration

Demonstrating the utility of the derived artefacts is not fully possible, since such action would require real world case study that reaches far beyond the scope of this thesis. However as Johannesson and Perjons point out, artefact demonstration is an important and necessary aspect of design science (Johannesson & Perjons, 2014). Through the use of STPA we have already demonstrated is applicability for improving system safety in social welfare administration. However we must yet still demonstrate how assumption-based leading indicators and the corresponding interventions can help make systems safer. To this purpose we will first build an assumption intervention matrix to illustrate, which intervention may be applicable to verify which assumption. Hereinafter, we will apply leading indicator program with the interventions to the control diagram of the Belastingdienst, seeking to answer the questions: ”Where in the control structure is the interventions best situated?”, ”When should the intervention be used?” and ”Who should execute and guide the intervention?”.

### 6.8.1. Assumptions Intervention Matrix

The purpose of this section is to showcase the process of determining how possible modes of checking, shaping and or hedging assumptions with the derived interventions. This exercise is done to verify that each assumption is meet with an intervention. In an ideal case with more than one, to provide the indicator program with the strongest possible indication for received signals. Further more the matrix can yield insight into which interventions may be regarded as most useful, through illustrating which interventions can be utilized for the most cases. The resulting 'range of applicability' provides insights reaching beyond the specifics of this case and therefore show cases the interventions utility for other, similar cases. Table 6.2 illustrates the matrix.
From the matrix we can gather that each intervention has numerous assumptions it is applicable for. Besides the ”hedging process” which is a mandatory intervention for each assumption stand out are: (1) DevSafOps team and (2) Audits. The value of STPA has been discussed previously and since an STPA analysis is also included in the ”hedging process” its utility and the corresponding necessity to utilize STPA for a system safety approach as been illustrated. Since this entire research approach is based upon an STPA is clear and does not further need to be outlined here as the standalone, most impactful intervention of the lot.

| ID | Assumption-based Lead Indicator | Audit | Dummy Data | Telematry | Red Team | STPA | DevSafOps | Batching | Hedging Proc. |
|---|---|---|---|---|---|---|---|---|---|
| 14 | Assumes that thrid party organisation follow provided data ontology | x | | | | x | x | | x |
| 3 | Assumes a 'normal' system workload & rate | x | | x | | | x | | x |
| 17 | Assumes enough development resources to adapt system | x | | | x | x | x | x | x |
| 16 | Assumes a safety orientated culture grounded in principals of Good Governance | x | x | | x | | | | x |
| 6 | Data quality in registration and models is assumet to be high | | x | x | x | | x | | x |
| 4 | Assums effective communication between sub-systems | x | | x | x | x | x | x | x |
| 1 | Assums it is feasable, System (not model, system) is abel to always classifiy individuals corretly | | x | x | x | x | x | | x |
| 12 | Assumes that there is a structural organisation facilitating this process of control | x | | | | | x | | x |
| 5 | Assums Mental model alignment throughout process and system | x | | | x | x | x | | x |
| 13 | Assumes that thrid party organisation provide accurate data | x | x | x | | | x | | x |
| 2 | Assums that misclassification in system is detectable | | x | | x | | x | x | x |
| 15 | Assumes colleted model data to be unbiase | x | | x | | x | x | | x |
| 8 | Assumes that system state is measurable | x | | x | | x | x | | x |
| 9 | Assumes that system data is interpretable | x | x | | | x | x | | x |
| 10 | Assumes that sufficant system state measurments are made | x | | x | x | x | x | | x |
| 7 | Assumes that incorporated feedback functions are executed | x | x | x | x | x | x | x | x |
| 17 | Assumes that the appeal process is functional and mitigates arising harms functions are executed | x | x | | x | x | x | | x |

Table 6.2: Assumption / Intervention Matrix

### 6.8.2. Demonstration of Intervention

Subsequently for each intervention we will show one possible mode of application within the previously illustrated STPA control diagram of the Belastingdienst. This is to illustrate each interventions applicability and functionality. However since the system depicted in the STPA does not longer exist in this specific form, the focus will directed towards demonstrating the interventions applicability, rather then to provide a full safety intervention for the system focused on alleviating all hazards previously identified. If one were to engage in such an analysis, the ”leading indicator monitoring program” would present a good point of departure. The assumption base lead indicator program will also be the point of departure for this demonstration, from were we will work our way through the different sub interventions. The full STPA including the interventions can be vied in appendix E.

Leading Indicator Monitoring Program

***Where?:*** The implementation of a leading indicator monitoring program should be directed towards a specific process a system is to execute rather then the sub-systems present in the system. For the case of the Belastingdienst this means that each of the divisions should implement a leading indicator program for their respective processes, not for the FSV system as a unit of measurement. The FSV system as a sub-system to the different process however should provide each respective leading indicator program with relevant information regarding their process and its own system state. Some of the information maybe general while other information may depend on the respective data relevant to the division. The leading indicator programs should stand in contact to each-other, since their is a likelihood of ripple effects if one program detects a failure. Moreover these programs have to follow a predetermined regiment for checking, shaping and hedging assumptions.

***When?:*** In the case of the Toezicht process of Toeslagen, the leading indicator program should receive relevant information from each sub-system/process, with which it is able to determine whether the made assumptions still hold true. Within the program times schedules regarding assumption checking have to be determined. For this purpose each assumption has to be fitted with checking schedule. Some of the assumptions maybe checked continouosly, while other assumption checks may only be triggered through the specific event, so called ”signposts” (N. Leveson, 2015). The table 6.3 below shows possible modes of scheduling such assumption checks for the assumptions determined for the Toezicht process.

| ID | Assumption-based Lead Indicator | Checking | Signpost |
|---|---|---|---|
| 1 | Assums it is feasable, System (not model, system) is abel to always classifiy individuals corretly | **Periodically** | Process/System Changes |
| 2 | Assums that misclassification in system is detectable | **Continuous & Periodically** | Scheduled Review |
| 3 | Assumes a ’normal’ system workload & rate | **Continuous** | |
| 4 | Assums effective communication between sub-systems | **Continuous & Periodically** | Scheduled Review |
| 5 | Assums Mental model alignment throughout process and system | **Periodically** | Process/System Changes |
| 6 | Data quality in registration and models is assumet to be high | **Continuous & Periodically** | Scheduled Review |
| 7 | Assumes that incooperated feedback functions are executed | **Continuous & Periodically** | Process/System Changes |
| 8 | Assumes that system state is measurable | **Continuous & Periodically** | Scheduled Review |
| 9 | Assumes that system data is interpretable | **Continuous & Periodically** | Process/System Changes |
| 10 | Assumes that sufficant system state measurments are made | **Periodically** | Process/System Changes |
| 11 | Assumes that there is a structural organisation facilitating this process of control | **Periodically** | Scheduled Review |
| 12 | Assumes that thrid party organisation provide accurate data | **Continuous & Periodically** | Process/System Changes,<br>Scheduled Review |
| 13 | Assumes that thrid party organisation follow provided data ontology | **Periodically** | Process/System Changes,<br>Scheduled Review |
| 14 | Assumes colleted model data to be unbiase | **Continuous & Periodically** | Scheduled Review |
| 15 | Assumes a safety orientated culture grounded in principals of Good Governance | **Periodically** | Scheduled Review |
| 16 | Assumes enough development resources to adapt system | **Periodically** | Scheduled Review |
| 17 | Assumes that the appeal process is functional and mitigates arising harms | **Periodically** | Scheduled Review |

Table 6.3: Scheduling Assumption-based leading indicator checks

From the previous section we can infer which interventions can be utilized to execute the respective assumption checks. Inadvertently, specific assumptions may be best utilize for different checks. Telematry can be utilized for a continuous monitoring, while audits are suited for scheduled reviews and a red team maybe planing ”attack experiments” after system changes to test resilience. Figure 6.6 shows the program implemented into the safety control structure of the Belastingdienst.

***Who?:*** From the figure 6.6 the importance of integrating the program with executing organisations becomes clear. The ”DevSafOps” team should be the program supervisor, since they retain a holistic perspective on the system. Direct management connection to the program should be maintained, to enable management to receive the valuable information about the system state gathered by the program. Simultaneously this close proximity should also result in the program retaining the necessary ”weight” to execute corrective measures once assumption failure has been detected or becomes more likely. Finally, the program needs to hold authority over the hedging process that is able to effect both development as well as operations, throughout the process chain. The ”hedging process” should be viewed as a significant event, once this process is triggered, trans-deivisonal cooperation to fix the issue and return to a safe system state is required. Therefore this process must deliver feedback to

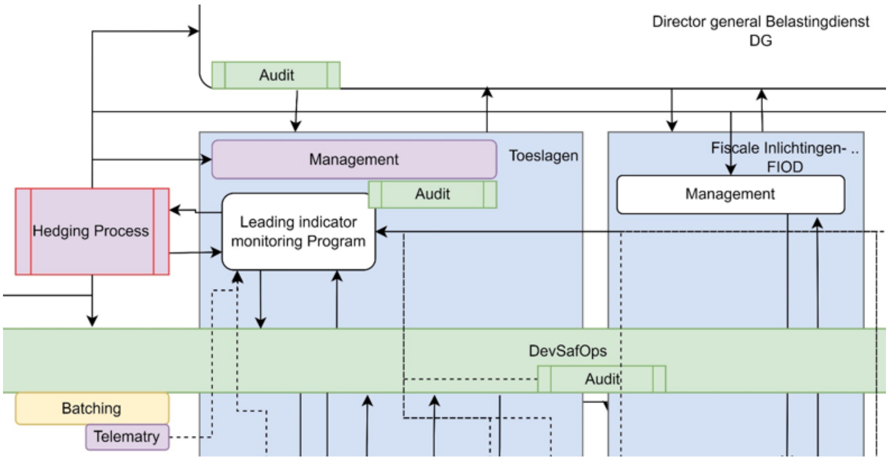


Figure 6.6: Assumptions-based Leading Indicator Program in hierarchical control structure

the director general of the Belatingdienst and should be incorporated into ever divisional management. Because some issue might only be resolved through adaption of workflows in different divisions, questions of authority and 'optimal resolution' may be raised. This is why executive supervision is needed, to ensure the divisional management groups are focused on conflict resolution rather then political positioning. Culture is crucial to this process and "blame games" should be avoided at any cost. For this purpose is might also be prudent to employ independent process consultants for this process. This will ensure transparency and raise accountability among the internal actors. For the Belastingdienst, the three pillars to a good intervention are hence, an extensive and responsive indicator program, a cross-functional "DevSafOps" team facilitating safety including leading indicators and management integrated hedging process focused on restoring system safety at all cost.

DevSafOps
The DevSafOps team illustrates a wide range of applicability. Its utility can be as part of shaping actions to prevent the failure of an assumption, by for instance contributing to high data quality in sub-systems. As well as in a checking capacity, through for example monitoring mental model alignment between different division and maintain process discipline. Furthermore DevSafOps teams can also take a pivotal the correct implementation of the leading indicator program itself. The program is in need of a interdiciplinary team that is able to derive the correct assumptions for operations made during development, monitor and update them, as well as deploy counter measures once an assumption is detected as 'failing'.

Audits
Audits are a valuable way of verifying many different assumptions. They great tool to utilize for assumption based leading indicator programs because they combine several benefits. Firstly, they are fairly easy to implement, within government structures, because audits are already in frequent use within public administrations. Secondly, they provide the added benefit of setting a 'standard', both viewed from a process as well as culture perspective. In accordance with this audits are also, to some degree, standardizeable. They do not directly depend on the underlying process or technology that is being audited, enabling to measure complex sub-systems/processes with a smaller number of key performance goals or indicators.

Telematry
Telematry can also be used as an intervention for a numeros assumptions. It is a great tool to record system state from which inferences regarding safety and the integrity of the assumptions can be made. As such it can naturally be used in a checking capacity. Telematry has the added benefit, that once it is set up within a system it provides continuous instant real world data. This ability makes it standout amount the different interventions, because its feedback is generated instantly. This ability is important due to the possible hazards that can arise because of "delayed control actions" (N. Leveson, 2018). Faster system state information can contribute to errors being detected earlier. In comparison to audits, which often span a time-frame of at least two weeks prior to reaching a conclusion, telematry can provide the the necessary early detection of assumption failure, that allows contour measures to mitigate all possible harms.

Red Teams
Read teams are also a vital component of a comprehensive intervention package, because they enable to detect weaknesses in the existing system structure or question the integrity of specific assumptions. They therefore

provide both checking and shaping capacity to the system. They are a great add on because the ”attracts” executed by the red team onto the system can test for several objectives at the same time. They can for example also be specified for special boundary cases, relating to both special system states or specific boundary groups within the global sample of cases.

### 6.8.3. Culture

That is an overall and very broad requirement, but absolutely vital to successful systems deployment and operations. If culture is not sensitized towards the correct values, many of the previously discussed interventions will faint without consequence. Important in this regard is to address that environmental conditions can effect culture, meaning the overall ’state’ of society has an impact on the effectiveness of is governance. Aside from such broad observations, policy itself evokes culture. Through policy politicians give critical impulses to the institutions in charge of executing them. Hence to prevent digital cage it is also vitally important to critically asses the underlying cultural implications of policy. The toeslagen case is a prime example for this as the recent report of the tweede kammer on the role of Parliament during the affair outlines. A fitting manifestation of politics being aware of this relation is the statement former secretary of state for finances Frans Weekers made during the parliamentary debate about stricter benefit fraud persecution after the ”Bulgarin fraud” case in 2014. He is said to have said: ”de goeden onder de kwaden zullen lijden” (the good ones among the bad ones shall suffer) prior to bowing to the parliamentary pressure to tighten welfare benefit control(“Toeslagenaffaire”, 2024). Consequently, the safety guarding culture must be implemented thought the entire system. This can be to some degree placed upon the responsibility of the DevSafOps team, but first and foremost must be highlighted, controlled and lived by management. Such an impulse hence should be given by the director general and is executive team, for the case of the Belastingdienst.

### 6.8.4. Continuous Monitoring and Evaluation

While the aforementioned design suggestions present a good point of departure to create effective safety interventions, they merely represent starting points to a ’safety journey’. To truly maintain an effective leading indicator program a comprehensive testing, monitoring and evaluation framework would need to be implemented. Each assumption should be assessed with qualitative and quantitative metrics, regular review cycles, and a mechanism for incorporating lessons learned into future safety planning through e.g. the use of STPA. This would allow the interventions to come full cycle and create a coherent and comprehensive safety plan that effectively addresses the constraints identified through your STPA analysis, ensuring the ADM system operates within safe boundaries.

## 6.9. Conclusion

This chapter addressed the third research question: *”How can safety control interventions for ADM systems in Dutch social welfare be structured, formalized, and applied in order to adhere to the identified objectives and requirements?”*. To this end, the chapter followed a two-stage design approach. Firstly, the chapter introduced a layered intervention approach that is based on an ’assumption-based leading indicator’ program, with respective interventions able to check, shape, and hedge the assumptions outlined as leading indicators. To delineate these interventions, each is explained by providing an answer to the following questions: ”*Why* is the intervention needed?”, ”*What* is the intervention?”, and ”*How* can the interventions solve the issue?”. Secondly, the intervention is demonstrated on the example of the Toeslagenaffair illustrating the intervention’s applicability, by providing an exemplification of how unsafe control actions and system hazards may be alleviated through their use. This was done by answering the questions of: ”*Where* in the control structure is the interventions best situated?”, ”*When* should the intervention be used?” and ”*Who* should execute and guide the intervention?”.

The design has applied the concept of ”assumption-based leading indicators” conceived by Leveson (N. Leveson, 2015), to the field of social welfare administration. As illustrated leading indicators have the potential to detect that the system is migrating to a ”higher state of risk”. This ability enables them to forecast the emergence of a digital cage and subsequently coordinate countermeasures to ensure this is prevented. Simultaneously, the assumption-based leading indicator program helps to improve system safety by facilitating active system-shaping mechanisms that ensure the system is constantly deployed and operated in a safe state. The different interventions all contribute to enforcement of the leading indicators in different ways. Through either checking, shaping or hedging the underlying assumptions. By dividing the design phase into to sections the thesis illustrates the concept underlying each intervention in a manner which allows the interventions to be generalized to other, relevant cases. Looking back to the derived system hazards a clear improvement of the systems ability to detect, measure, interpreter and recover system states can be deduced for the toeslagenaffair. The main conclusion of this chapter is the implicitly demonstrated potential of assumption-based leading indicator programs as a potent weapon against unwanted emergence in complex systems. Specifically, also in cases were ADM system elements obscure system traceability. The following chapter will continue with the evaluation of the designed intervention and provide suggestions for further evaluation.

# 7 Design evaluation

This chapter covers the fifth step of the design science process and considers the fourth research sub-question: ”What is the utility of the safe control interventions for ADM systems in Dutch social welfare?”. To answer this question the following section will first provide some general information on the evaluation approach, before moving on to present the workshop evaluation and the corresponding results. This will be followed by a reflection, as well as suggestions for further evaluation.

## 7.1. General Evaluation Information and Approach

The conducted evaluation represents an ex ante evaluation of the design (Johannesson & Perjons, 2014) and considers the derived interventions in their utility for a wider area of application than only within the analyzed Toeslagenaffair context. An ex-ante evaluation represents the only feasible evaluation approach to the interventions at this time since a direct implementation into the Toeslagensystem is no longer possible. Furthermore, ex post evaluation would represent a long-term study that reaches far beyond the scope of this thesis. The ex ante evaluation has the added benefit that it provides a quick and inexpensive approach to obtain feedback on the design (Johannesson & Perjons, 2014). However, we note that ex ante evaluations can result in false positives, and therefore outcomes from this evaluation should rather be viewed as a formative evaluation to inform future work rather than reliable results for summative evaluation of the utility of the interventions for social welfare(Johannesson & Perjons, 2014).

As described in chapter 3, the main evaluation method for this design science research was a planned workshop series. Figure 3.2 illustrates the workshop process. For the workshop, we orientated ourselves on the guidelines provided by Thoring et al. (Thoring et al., 2020). Laying a particular focus on ”Workshop Outcome Quality” we utilized several of the methods suggested by Thoring et al. The method chosen for artifact analysis was the ”parking lot method” (Miro, n.d.). This method essentially asks the experts to differentiate with the help of two important aspects: (1) Ease of Implementation and (2) Potential Impact and hence provides a good indication of the utility of the respective intervention for the field of ADM in social welfare administration. Therefore the workshop was set to a one-hour time frame. The sessions were planned to be conducted online with the help of Microsoft Teams and Miro. Participants should first be introduced to the topic through a presentation of the thesis and the respective interventions. This was planned to take approximately 20 minutes. Hereinafter the focus was to be shifted towards the Miro board which the participants should then be instructed to use to categorize the different interventions into the parking lot grid. The designed Miro board template can be viewed in figure 7.1. Each participant was to receive their own colored sticky notes which they would able to drag and drop to the respective field they feel is most suitable for the intervention in question. During this process, participants were able to ask questions as well as engage in an open discussion about the interventions. The respective choices were however made independently of one another. Subsequently, a discussion about their respective choices was to be conducted for approximately 15 minutes. Finally, the workshops ended with a discussion of the concept of ”assumption-based leading indicators”. This approach enabled us to utilize video recordings, observation & notes, a group discussion as well as artifact analysis as methods to increase outcome quality.

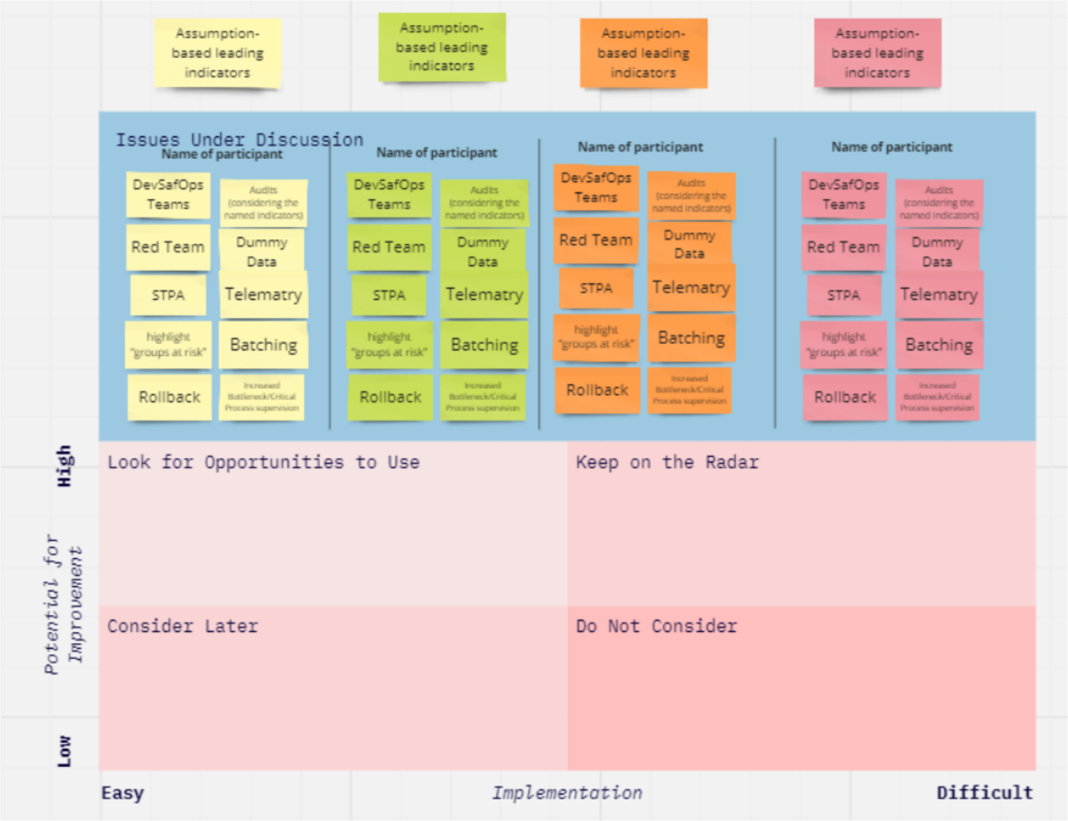

Assumption-based leading indicators
Assumption-based leading indicators
Assumption-based leading indicators
Assumption-based leading indicators
Issues Under Discussion
Name of participant
Name of participant
Name of participant
Name of participant
DevSafOps Teams
Audits (considering the named indicators)
Red Team
Dummy Data
STPA
Telematry
highlight "groups at risk"
Batching
Rollback
Increased Bottleneck/Critical Process supervision
Look for Opportunities to Use
Keep on the Radar
Consider Later
Do Not Consider
High
Potential for Improvement
Low
Easy
Implementation
Difficult

Figure 7.1: Workshop: parking lot template Miro board

## 7.2. Results

The workshop evaluation was set to be a workshop series with each meeting consisting of approximately 2-6 workshop participants. In total 16 individuals were asked to participate and provided with a time frame of 3 weeks to find possible time slots for a workshop. Due to the incompatibility of time frames and several unfortunate incidents of illness and late scheduling conflicts, only 2 experts could participate. This is quite an unfortunate outcome regarding the evaluation that will be further discussed in section 7.3. Moreover, these individuals shared the same background stemming from research. Regrettably, experts closer to the Dutch public administration were not able to join.

The workshop was conducted within a one-hour time frame. The session was held online with the help of Microsoft Teams and Miro. Participants were first introduced to the topic through a presentation of the thesis and the respective interventions. This process took approximately 20 minutes. Hereinafter the focus was shifted towards the Miro board which the participants used to categorize the different interventions into the parking lot grid. Subsequently, a discussion about their respective choices was held. Finally, the workshop ended with a discussion of the concept of ”assumption-based leading indicators”.

### 7.2.1. Interventions

The results of the parking lot analysis can be viewed in figure 7.2. From this analysis, we can see that the interventions have been evaluated similarly by the two participants indicating their shared background and understanding of the issue at hand. Furthermore, we are able to observe that neither of the participants deemed the interventions as un-useful, indicated by the parking lot ”do not consider” staying empty. Within this evaluation, the interventions making up the ”hedging process”, namely ’highlight groups at risk”’, ’roll back’ and ’STPA’, have been put to the participants separately. However, the process has been explained during the workshop presentation. For purposes of a more detailed perspective on the sub-components, the process was split.

From the gathered evaluation we can see that the interventions can be roughly split into two categories, either viewed as ”higher potential, but difficult to implement” or ”easy to implement, yet lower potential”. One can observe that the split between these two groups can be made quite intuitively between more ’technical’ interventions and ’organizational’ interventions. With this comparison, we see that technical interventions are assessed as being relatively easy to implement while organizational interventions are more difficult to implement, yet also provide a larger potential for improvement. ”DevSafOps” teams that have been previously identified as pivotal are rated as the most difficult to implement. This correlates with the comments made in the design demonstration section 6.6. As pointed out there, creating such cross-functional teams through multiple divisions within a government organization is likely to require high investment in human resources. During the discussion it became clear that the ’technical’ interventions were perceived to only be able to apply to ’technical’ system components, this however is not necessarily the case. As pointed out with telemetry, rollbacks, or dummy data previously these interventions could also be applied to human decision-makers within the system at points of decision-making.

### 7.2.2. Assumption-based leading indicators

The assumption-based leading indicators have been categorized as a relatively feasible intervention with the potential to improve the system.

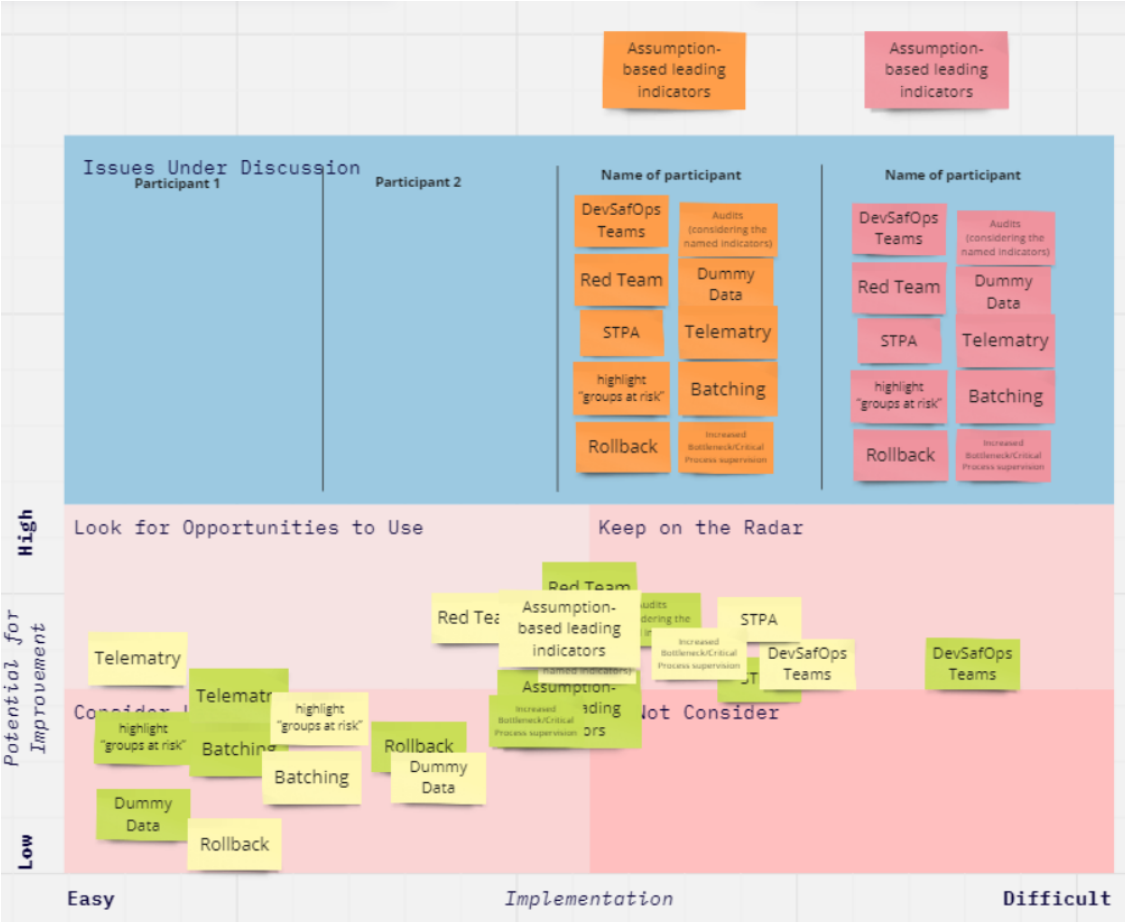

Assumption-based leading indicators
Assumption-based leading indicators
Issues Under Discussion
Participant 1
Participant 2
Name of participant
DevSafOps Teams
Audits (considering the named indicators)
Red Team
Dummy Data
STPA
Telematry
highlight "groups at risk"
Batching
Rollback
Increased Bottleneck/Critical Process supervision
Name of participant
DevSafOps Teams
Audits (considering the named indicators)
Red Team
Dummy Data
STPA
Telematry
highlight "groups at risk"
Batching
Rollback
Increased Bottleneck/Critical Process supervision
Look for Opportunities to Use
Keep on the Radar
Consider
Not Consider
Red Team
Red Team
Assumption-based leading indicators
STPA
DevSafOps Teams
DevSafOps Teams
Telematry
Telematry
highlight "groups at risk"
highlight "groups at risk"
Batching
Batching
Rollback
Dummy Data
Dummy Data
Rollback
High
Potential for Improvement
Low
Easy
Implementation
Difficult

Figure 7.2: Parking lot method evaluation outcome

## 7.3. Reflection

The utilized evaluation method within this thesis has several weaknesses that should be brought to the reader’s attention. This section reflects on several of these aspects both analyzing the general flaws of the evaluation and reflecting on the choices made by participants the the procedure used for the evaluation.

### 7.3.1. General aspects of the Evaluation

Workshop

As previously mentioned the workshop was initially planned as a workshop series, however, due to lack of decisive interest workshop was ultimately only able to attract two participants. Numerous reasons can be identified to have attributed to this outcome. While not being able to identify the right ’audience’ for the workshop is likely to be a large contribution, it might also point towards a lack of awareness of the importance of system safety thinking for this issue. This can not be proven conclusively, however, indicators at several stages of the research hint towards this fact. Indicators such as the amount of available literature, the number of experts available for discussion, and the number of individuals who expressed disinterest in participating in the evaluation. The meager participation ultimately can said to have had an impact on the evaluation. Since 2 individual to evaluate a design, though be it their experts, are not representative of building strong inferences regarding design utility, etc. In addition to this aspect, the respective time limitation of a 1-hour workshop further diminishes the output quality of the evaluation. The topic and design at hand are rather complex and require some introduction prior to further evaluation. This is a trade-off that was made in order to accommodate more potential participants, which in turn would transfer to more output. Unfortunately, this calculation did not materialize. With more time at hand, the respective interventions could have been scrutinized in more depth possibly even making time to evaluate the respective assumptions derived for the Belastingdienst system. A more extensive evaluation could have contributed to distilling the respective potential lying in each intervention further, to subsequently provide a better ground for recommendations regarding which intervention to follow up on with further research. Moreover, a full workshop series could have provided insight into the different perspectives different stakeholder groups surrounding the Dutch social welfare system have on the issue and their respective ideas on how to solve them. Such information would have been similarly valuable since it could have been used to cater the recommendations to the respective expectations of the stakeholder through engaging in their respective ”language”. A contextualized demonstration of each intervention presented during the workshop could also have improved participants’ understanding of the respective concepts. This however would have required more time and a specialized approach for each group of participants, since they do not necessarily share the same background different instances/examples might have been needed for different participants.

Finally, we can deduce that the conducted workshop lacks robustness in several key matrices. Nonetheless, viewed as an ex ante evaluation the workshop provides some interesting indications that can be investigated in future research.

Connection to System Safety Constraints

Besides the criticism that can be made regarding the chosen evaluation technique one might add that the approved presented evaluation does not address the previously derived system safety constraints as design requirements. In a full evaluation the design has to be evaluated against every design requirement (Johannesson & Perjons, 2014), this was only partially done here. However on the contrary we can point towards the ex ante nature of the evaluation for this thesis, hence the evaluation of the interventions was also applied during the design face, while iterating over useful ”leading indicators” and the respective checking, shaping, and hedging interventions. Consequently, each of the requirements has found consideration in the design output. Furthermore, the demonstration section elaborated on how the specific interventions could alleviate the identified unsafe control actions and system hazards for the Belastingdienst toeslagen case. This demonstration, in part therefore also represents an evaluation, illustrating that the interventions can indeed help to fulfill the safety constraints previously determined as design requirements. Readers can further confirm this by comparing the two hierarchical control structures made of the system in appendix C and E respectively.

### 7.3.2. Reflection on Participants

Several of the observations made in the result section have to be contextualized regarding the individuals who participated in the workshop. In the result section we observed a clear tendency of the participants to favor organizational interventions before technical interventions. Similarly, we also observed that the organizational interventions in the eyes of the participants would be hard to implement. However, these observations are not deterministic facts but rather dependent on the experience, tendencies, and understanding of the ”problem area” of the respective participants. Consequently, a different outcome is likely to have materialized if other individuals with another background had been posed with the same question. This clearly highlighted the subjective characteristics of the evaluation results and has to be considered in order to derive relevant, scientific conclusions from the research.

The subjective nature of the evaluation produces several conclusions. Firstly, to distill the true utility of each intervention and the leading indicators, empirical studies have to be undertaken. Such studies should focus on evaluating singular interventions in several different contexts since the environmental setting will likely have a large influence on the respective utility of the intervention as well. Secondly, as different groups will always evaluate the utility of the interventions differently, the evaluation illustrates a clear need for a more holistic problem-solving approach within this field of study. Through the STPA analysis, we can deduce that several issues representing different types of problems contribute to the emergence of digital cages. We have illustrated technical, organizational, and institutional issues that perpetuate the emergence of digital cages. Consequently, we can deduce that neither narrow intervention approach stemming out of one of these areas will be successful in eliminating the digital cages phenomenon. In contrast, inter-disciplinary cooperation is needed to form a holistic intervention approach that is able to coordinate the different afford undertaken in each of the respective sub-disciplines to combat the

problem.
Within the evaluated subjective perspective of the workshop participants, and similarly in a large number of previously analyzed scientific literature a tendency to focus on the ”blame of the technology” is present. In the workshop, this perspective manifests itself in the participant’s tendency to reject technological interventions, while in some of the analyzed literature, this perspective is illustrated through their focus on the ”natural characteristics” of the technology that perpetuates the problem. These positions are based on valid reasoning and clearly outline part of the issue. The impact of the utilized technology on the problem space has also been observed and criticized within this research. However, we must also point out that within this research other reasons for the emergence of digital cages have also been made apparent. In many of these cases, Levesons work has provided us with clear delineations of how the respective issues may perpetuate safety issues within complex systems, see (N. Leveson, 2011). In our research organisational inadequacies that caused the toeslagenaffair are clearly visible.

From this, we can further strengthen our argument for a holistic problem approach to the digital cage problem. Stakeholders have to be aided in realizing that both organizational and technical inadequacies contribute to the formation of unsafe system states and have to be combated holistically to ensure system safety. The practice of singling out specific issues within this problem area as most perpetrators neglect the lessons from systems and systems safety thinking presented in this research. As we have shown emergence is a phenomenon that arises through the interactions between subsystems. These include technological, social as well as institutional components.

### 7.3.3. Reflection on own Procedure

Just as the respective experiences of each of the evaluation participants are likely to have influenced the results, so two the following evaluation procedure will have influenced the evaluation results. We have already pointed towards general issues with the evaluation process, such as time scarcity. Aside from this, the choice of how to present the problem and each intervention will have influenced participants in their perception of the problem and the corresponding potential for improvement that can be attributed to each intervention. An example of this is the intervention ”Dummy data” which could have be phrased differently to highlight its utility for not only testing adm systems but also as a quality control mechanism for human sub-processes in the system. ”Dummy data” may have evoked a clear connection already present in the participant’s mind. Such as ’testing data’ for machine learning models. This became apparent in the subsequent discussion with the participants, where ”dummy data” was understood as a process to check the algorithms present in the system not however as an intervention to verify case-worker decision-making or to trace the ”business logic” through the number of sub-organisations.

This relation points towards the need for empirical evaluation. Through an ex post evaluation of the designed interventions and the assumption-based leading indicators, such correlations are more likely to be avoided. Since the actual implementation of such interventions would help to eliminate disparities in the ”mental models” of different stakeholders in their understanding of the intervention. However, even in such a setting, this causality can not be avoided entirely, because the respective ’attitude’ stakeholders exhibit towards the respective interventions is likely to have an impact on the intervention’s effectiveness and correspondingly the ”quantitative” data that can be collected from such an evaluation. The latter argument highlights the need for an inclusive problem approach that utilizes the right ”language” to address each stakeholder taking needs, wants, and experience into account.

## 7.4. Further Evaluation

While the afore-presented evaluation of the interventions presents a point of departure to determine their respective applicability and utility, it is not comprehensive in nature. To truly assess these characteristics the interventions each need further investigation and testing. To this purpose, empirical studies should be commenced each addressing a clearly defined subsection of the intervention presented in this research.
Specifically, the assumption-based leading indicators should be evaluated further in their utility to circumvent the issue of digital cages in social welfare. The indicators are the basis on which the design intervention hinges therefore their relevance has to be verified through further research. This research should focus on deriving qualitative output that outlines clearly, what specific assumption can be used within an assumption-based leading indicator program for maximum utility. As an example, one could research the leading indicator utility of ”workload”. Such research could entail discrete event simulation that models the workflow through the system including respective capacity etc. From such models inferences regarding the effects of variance in workload me built. Ultimately, such research could evaluate the applicability of specific leading indicators and give practitioners guidance on how to utilize them in their respective systems. Other indicators could be verified through field studies, such as mental model alignment.
Aside from the leading indicator program, further evaluation has to be directed towards the ”hedging process” outlined in this research. Its feasibility is likely to be highly dependent on the respective system context, non the less the concept of ”system rollbacks” to mitigate the emergence of harm merits such investigation, as it would greatly

improve the mitigation ability present in social welfare. Clearly, such capabilities are pivotal, since the ultimate goal should be to mitigate harms converging onto citizens.

## 7.5. Conclusion

This chapter addressed the last sub-research question by outlining the evaluation method used and subsequently analyzing the results achieved by this evaluation process. The conducted evaluation points towards the potential of organizational interventions for safety improvement. Additionally, it also points toward the need for a more holistic problem approach that enables stakeholders to combat the issue of digital cages in an interdisciplinary fashion. The evaluation reflection outlines some clear weaknesses posed by the evaluation and emphasizes the need for further evaluation, specifically the need for empirical studies that can provide quantitative proof regarding the utility of the proposed interventions. This especially applies to the assumption-based leading indicators for ADM systems in complex social welfare system of systems. Finally, the evaluation also points toward a lack of awareness regarding the importance of system safety thinking for modern administrative systems. The following chapter will conclude the thesis with a discussion of selected aspects of this research, the conclusion, and a recommendation for feature research.

# 8 Discussion & Conclusion

This chapter will finalize this thesis by firstly discussing relevant concepts and findings presented in this research, secondly positioning this research in the existing body of literature by outlining its research contributions, giving stakeholder recommendations based on the findings this research has made, and finally discussing suggestions for future research subjects corresponding to the conclusions of this thesis.

## 8.1. Discussion & Reflection

Within this section, we will reflect on the core concepts and methods presented in this research. Subsequently, the respective concepts will be discussed paying a particular emphasis on lessons learned and reflecting on the limitations of these concepts. The section will be finalized with a discussion about the implications that the designed artifact evokes in regard to future public administration management.

### 8.1.1. Digital Cages

The digital cage concept as introduced by Peeters and Widlack (Peeters & Widlak, 2018) was one of the pillars central to this research. Its delineation of the relationship between information infrastructure, bureaucracy, and citizens subject to cybernetic control motivated this research. Weber's ”iron cage” builds the conceptual foundation of the modern ”digital cage”. The concept of the ”iron cage,” originally coined by Max Weber as *”stalwarts Gehäuse”* (shell as hard as steel), is one of the most influential and enduring ideas in the history of social thought (Baehr, 2001). Weber introduced this metaphor in his seminal work on the rise of bureaucratic capitalism. The *”stalwarts Gehäuse”* represents the process of rationalization and bureaucratization that, according to Weber, increasingly characterized Western society, particularly in the wake of the Industrial Revolution. The Weberian perspective on bureaucracy however is broad, Weber viewed bureaucracy as ”living machines”, because it shares the same *”congealed spirit”* as lifeless machines (Baehr, 2001). As such the phenomenon is transferable to artifacts sharing the similar *”congealed spirit”* (Baehr, 2001) and thus provides a great metaphor to describe the ”robotic” nature of ADM-guided decision processes in modern social welfare administration.

Interestingly, both in the digital age research and system safety, constraints play an instrumental role. Within Weber's research, the constraints restricting the individual freedom of citizens are highlighted, while in system safety research constraints regulating the behavior of the system are emphasized. This highlights the duality of constraints in system theory. Ultimately, human interpretation is needed to determine whether a constraint is harmful or beneficial. Here we are drawn back towards fundamental discussions currently taking place in the scientific community surrounding contemporary governance, technology, and its impact on discretion, see (Bullock, 2019) also (Bovens & Zouridis, 2002). With a shift from street-level to system-level bureaucracy, means by which discretion can be exercised have also changed. Ultimately, we believe that 'discretion' historically can be viewed as an arbitrary tool to avert consequences resulting from institutional biases away from citizens. Rigid IT infrastructures that are implemented in adaptive complex systems inhibit this vital control mechanism and thereby increase systemic risks.

There is a potential pathway to reintroduce discretion into these highly automated systems, through system safety, especially with the integration of control actions. Control actions in system safety are designed to manage risk and ensure safe outcomes by influencing system behavior either through direct intervention or by modifying constraints. Applying this thinking to modern public social welfare, control actions could be implemented not just for safety but also for ensuring that systems retain a level of human oversight and discretionary power. This can be achieved by designing systems that require human intervention in critical decision-making processes, or by setting thresholds that trigger a review when automated decisions fall outside of established normative ranges.

Introducing such measures could help maintain the balance between efficiency and personalization that the bureaucracy strives for but too often gets wrong. By embedding discretionary ”signal posts” (N. Leveson, 2018) within the digital age, systems can be made more adaptable to changing circumstances. This approach respects the intent of social welfare systems to serve the public efficiently without sacrificing the individual care and consideration that discretion allows. Thus, system safety principles, particularly through the thoughtful application of control actions, offer a compelling framework for retaining human judgment in the increasingly automated fabrics of system-level bureaucracies.

### 8.1.2. STPA for ADMs in social welfare

As illustrated within this research the methods entailed within system safety provide a largely unrealized potential to improve system safety for the modern social welfare system. Especially STPA enables designers to capture complex causes and interdependencies between sub-systems that can subsequently be addressed through improvements in the systems control structure. Yet, as this thesis illustrates applying STPA to this subject is no straightforward feed. As illustrated through the design output this thesis has produced the implications created by STPA, if applied to such a complex system, are numerous. This results in a broad number of different problem areas, that are difficult to address within a narrow intervention framework. Instead, we see that there are a multitude of interventions needed to address the problems identified by STPA. While the holistic perspective of STPA provides a great deal of benefits, see (N. Leveson, 2011), in the case of complex social welfare systems it also has the drawback that it does not provide a narrow set of actionable interventions that practitioners can execute. This diminishes the likelihood that practitioners are willing to implement the suggested system constraints, simply because the necessary changes to the system are too extensive and complicated. The holistic scope of STPA also contributed to the respective interventions lacking specificity and rigor. Since the number of interventions resulting from STPA were so high, none of them were addressable in a rigorous enough fashion to truly illustrate their respective utility. If applied in practice STPA could run into similar problems if not applied with sufficient rigor, understanding, and necessary authority. A clear focus has to be given on what part of the system should be improved. Nonetheless, the utility of STPA for identifying system hazards in complex social welfare systems is demonstrated through this research. Consequently, they should be utilized in social administration, however, safety designers need to consciously focus on providing specific interventions that can help to improve the system safety.

Lessons learned from applying STPA

The process of applying STPA to the subject of social welfare administration has yielded some insights specific to these types of systems, which will be discussed hereinafter. As previously mentioned the utilization of STPA in this subject area has not been done before. Since STPA is traditionally applied to other subject areas with different characteristics, the underlying methodology is at times difficult to follow for this respective problem area.

For one determining the system boundaries, an important step within the STPA analysis (N. Leveson, 2018), is hampered due to the complexity of modern public governance. As mentioned in chapter 1, the social welfare system is a complex system of systems, this system is further integrated into other governmental functions difficult to dissect in their entirety. This makes it very difficult to clearly outline system boundaries, even in single organizations, as the case of the Belastingdienst exemplifies. This instance emphasizes the importance of interfaces. Meaning because the system boundaries are somewhat more ambiguous and dependent on the observer’s perspective, the observer has to direct extra attention toward the ”input” and ”output” the determined system receives and generates. This includes environmental conditions as well as interactions with other systems adjacent to the unit of analysis.

Another issue previously eluded to is that the social welfare system does not control artifacts subject to physical, or ”passive” control actions, but instead controls mostly ”cognitive processes” (Jakubiec, 2022). This has the consequence, that it is more difficult to identify mechanisms of control and feedback present in the system because they are not standardized in the same fashion as this would be the case in traditional areas of application for STPA such as aerospace. And secondly, the ’curse of flexibility’ (N. Leveson, 2011) is magnified. This flexibility makes it difficult to conceive future ”loss scenarios” and therefore complicates also the task of identifying system hazards. Additionally, the ’migration’ towards a higher level of risk (Rasmussen, 1997) is harder to detect, as there are no physical indicators preset in the system that could be used for verification. Information, however, especially stemming from complex socio-technical interactions, can retain ambiguity. Consequently, safety controllers should pay equal attention to each cause of unsafe control action described in chapter 4.3: inadequate control, feedback, or data/information. We want to highlight data and information in this respect because in comparison to traditional applications of STPA data in social welfare can be more ambiguous and often is not structured through a uniform ontology.

In comparison to the traditional application of STPA, the hierarchical structure in the system is also more ambiguous. The social welfare system is characterized by a lot of cross-organizational cooperation that sees sub-organisations meet on eye level and cooperate to reach their respective objectives, rather than a structure of command and control. Processes are highly impacted by sub-organizations the respective process owner only has limited control over. This can lead to difficulties in visual representation because the control structure grows in ”width” rather than ”depth”. Moreover, it makes the representation of control, feedback, and information flow multi-dimensional. Additionally, several organizations are utilized for different processes that pose different requirements, which can not easily be prioritized. This can lead to identified hazards for one process which are however caused by characteristics necessary for other processes. These tradeoffs exist between different sub-processes and are very difficult to visualize within one hierarchical control structure. Conversely, the complexity of mental model alignment between the sub-organizations is oversimplified through the existing control structure, as it only shows a one-dimensional (process) perspective of the system.

### 8.1.3. Assumption-based leading indicators

The concept of assumption-based leading indicators introduced in this research for ADM use in social welfare administration represents a core result of this thesis.

An Assumption-Based Leading Indicator program is a proactive strategy designed to enhance system safety and foster a culture of safety within organizations by focusing on the foundational assumptions that underpin all system designs and operational decisions (N. Leveson, 2015). This approach moves beyond traditional metrics that typically register problems only after they have occurred. Instead, it aims to identify and address potential failures and safety risks at an early stage, thereby preventing incidents rather than merely reacting to them. The essence of a program lies in its emphasis on assumptions of often unexamined premises that are considered given within the design context of a system (N. Leveson, 2015). These assumptions can relate to various facets such as the reliability of technology, the predictability of human behavior, regulatory standards, etc.. The critical task is identifying these assumptions clearly and understanding their pivotal role in the functioning of the entire system.

The proactive nature of an assumption-based leading indicator program is especially valuable in complex systems where failures can have unforeseeable consequences. By monitoring these leading indicators, organizations can intervene early, often before any actual failure occurs. This allows for adjustments to be made in operational practices, system designs, or safety protocols, thus maintaining system integrity and safety. Moreover, their implementation also contributes significantly to the development of a safety culture within organizations. When employees and management actively engage in identifying and monitoring assumptions, there is a shift towards a more safety-conscious mindset. This culture is characterized by continuous improvement to adapt based on the feedback provided by the leading indicators. It encourages openness in discussing potential problems and fosters a learning environment where safety becomes a shared responsibility (N. Leveson, 2018). However, regular updates and reviews of the assumptions and their corresponding indicators are essential. As systems evolve and external conditions change, previously valid assumptions may become outdated. Regularly revisiting these assumptions ensures that the system’s safety measures remain robust and relevant. Additionally identifying meaningful assumptions can be a highly specific and difficult task, as the applicability of general indicators has not proven to be fruitful in the past, see (N. Leveson, 2015). The in this thesis provided assumption-based leading indicators therefore should not be taken at face value but rather used as a starting point to identify and refine other more meaningful indicators to the specific system in question. Given this complication more research into ADM relevant leading indicators is necessary. Especially in the domain of social welfare administration.

In conclusion, an assumption-based leading indicator program offers a dynamic and effective approach to improving system safety. By focusing on the underlying assumptions of system operations and designing indicators to monitor these, organizations can anticipate and mitigate risks proactively. This not only enhances the safety of the systems but also ingrains a culture of safety across the organization, making safety an integral part of the operational process rather than an afterthought (N. Leveson, 2015). Such a strategic approach is indispensable in managing complex systems where the cost of failure is high.

### 8.1.4. Implications of the Design Artefact

Several implications can be drawn from the design artifact.

Designer Overload

Firstly, as mentioned previously, the artifact proposed is quite extensive in nature. While there are many positive aspects to adopting a system safety perspective for social welfare administration, this also implies more requirements for system designers. This creates a danger of ’overloading’ the designers with too many requirements, something we highlighted previously, in the context of deployment size, as a negative contributing factor. Overloading designers and thereby inadvertently the design process bear the risk that the necessary requirements can not be effectively implemented into the system. It is therefore important that system safety is implemented in a way in which it informs the design process in a meaningful way and aids designers in their tasks rather than to impose further administrative tasks onto them, something that in practice is often associated with compliance. Tools such as STPA mustn't be utilized in an ex post fashion, directed towards proofing compliance, but rather as a critical designer assessment tool to improve system safety. As such STPA and system safety, just like the system itself, should always be viewed as a ”work in progress” that can continuously be improved. Moreover, not overloading system designers can be made a leading indicator, thereby bringing the utility of STPA and system safety full circle.

#### Incentive structure and Sanctions

Another implication of the design artifact is the importance of achieving the right incentive structure, for system designers and operators to act in accordance with the safety constraints determined for the system. We outlined previously how the contemporary project contracts with third-party developers perpetuate a short-term, ”get it over the line”, attitude within the initial project phase. Similarly, we pointed towards policy as a building block for an organizational culture that incentivized a stricter, unsafer operational approach. Revisiting these incentive structures to form a better environment for safety is important. One aspect that has been overlooked up to this point by this thesis is the importance of sanctions in facilitating this process. Sanctions can help shift the incentive or rather constraint structure, thereby contributing to a safer system state.

One aspect through which this can be achieved is to hold system developers more accountable. This could be achieved through mutating the contract terms with third-party developers. Adopting a broader goal for ”system handoffs” that does not merely focus on proving the technical functions of the system but also highlights the safety performance of the system and its outcome. Additionally, third-party developers should face some of the liabilities resulting from the misclassification of citizens, helping to ensure that their interests are incorporated into the design process rather than only the requirements proposed by the client organization.

Another sanction worth exploring is the sanctioning of the state though creating a legal obligation for compensation if harms were created. This sanction could work in two ways. Firstly, by incentivizing system operators to be more careful due to the implied, quantifiable financial risk of misclassification and harm. Repurposing the ”congealed spirit” (Weber & Tawney, 1930) of bureaucracy to the advantage of safety. Moreover, shifting the ”burden of proof”, towards the government in cases where digital cages may have formed. The latter can significantly help citizens who are subject to a digital cage, by lightening the ”pressure” they are subject to while fighting for their rights.

#### Public Management

Another implication of the design artifact presented is the need to adopt a different approach to public management. The STPA analysis illustrated that while the technical interfaces between different organizational silos and administrative bodies are continuously being broken down, the same can not be said for cross-functional cooperation. This is an area that must be improved through management initiatives that help form more inclusive communication cycles. Additionally, with the increased adoption of ADMs in public administration, managers must recognize the shift from an operational to a controlling, monitoring function in their department. Emphasis should be put on qualitative rather than quantitative output. This can help mitigate hazards evoked through the metaphor of Weber's expression, the ”congealed spirit”. The stronger focus on qualitative metricizes and a clearer ordinance to focus on system control can also help to prevent ”ripple effects” and conversely break the emergence of a digital cage. Applying lean logic, if stronger control is implemented within each working center, the likelihood of passing misclassified data to downstream workstations is reduced, reducing the overall likelihood of ripple effects emerging in the system landscape (Kim et al., 2016b).

#### Data Quality

One aspect closely related to the previous is data quality. Through the STPA analysis of the Toeslagenaffair, it became apparent that at several key process steps, ambiguity about data and its quality was inadvertently accepted. It is therefore prudent to advise system designers and operators to adopt a more rigorous approach towards data ontology, these should be determined, documented, and communicated with all relevant stakeholders to circumvent misalignment and inadvertent misuse. Setting a consistent data ontology can enable system operators to avoid many of the causes for unsafe control action, as these were often related to the inadequacy of provided feedback or information/data, see chapter 4. As mentioned before this is also one of the areas where public administration diverse from the usual area of application for STPA. In aerospace, for example, communication is rigorously structured to avoid miscommunication, this also includes that the ontology of the transmitted data, and information is clearly defined. For the area of public administration, this issue comes back to the issue of mental model alignment, where more effort has to be put into a holistic process alignment across different organizational silos. In the case of data quality, this should be a task facilitated by the role of already existing data officers. The role of data officers should be expanded to not only handle GDPR problems but also data quality issues where the existing data quality does not align with previously set quality standards for the respective process. Internal data wikis could help facilitate this alignment.

Culture & Policy

We have already addressed that in light of the increased adoption of ADM systems in public administration, public management must change. This also applies to cultural aspects of public administration. Leveson points out culture as one the most impactful factors determining long-term system safety, the right culture can improve effectiveness and safety (N. Leveson, 2011). Similarly, the socio-technical analysis conducted in chapter 4.1, illustrates how cultural aspects of governance contributed to the emergence of the digital cage in the Toeslagenaffair. From this analysis we can get a sense of the organizational pressure exerted onto the Belastingdiesnt and carried on within the organization itself, to crack down on fraud, through stricter decision-making. This culture conversely neglected safety concerns regarding misclassification and the there out resulting administrative exclusion. As Frans Weekers remarked during the ”Bulgarian fraud” case in 2014: ”de goeden onder de kwaden zullen lijden” (the good ones among the bad ones shall suffer) prior to bowing to the parliamentary pressure to tighten welfare benefit control(“Toeslagenaffaire”, 2024). This statement implies that decision-makers at the time were aware of the impact the policy and the thereout resulting organizational, and cultural change would have on the situation. In the STPA handbook, Leveson identifies several cultural characteristics detrimental to a safety culture (N. Leveson, 2018):

- Culture of Risk Acceptance: This culture is based on the assumption some accidents, or incidents are inevitable. They are considered to be part of the tradeoff function of productivity and efficiency. This assumption is often accompanied by the belief that accidents result from a lack of responsible behavior on the part of individuals. This opinion believes that if individuals and groups would act responsibly and safely, accidents would be reduced, thereby neglecting safety as a function of system design. Overlooking that safety can be achieved by design, and continuously be improved through proper control mechanisms.
- Culture of Denial: In a culture of denial credible risks and warnings are dismissed without appropriate investigation. Responsible decision-makers are not interested in listening to problems, they only want to hear good news, so that is what they are told. The effort is directed towards proofing that the system is acceptably safe, not identifying the ways it might be unsafe. The cycle of confirmation bias is born.
- Culture of Compliance: The underlying cultural belief is that complying with regulations will lead to acceptable results. Because regulatory agencies tend to focus on certifying that a product or service is safe, post-fact assurance is emphasized, and often extensive “safety case” arguments are produced with little or no impact on the actual system or process safety. Compliance is practiced for compliance’s sake.
- Paperwork Culture: Practitioners of this culture believe that lots of documentation and report paperwork results in system and process safety. Large amounts of documentation are produced with little real utility for design and operations. Safety-related documentation may be produced by a group that is independent of and has little interaction with those who are designing and operating the processes and system, deployment, or operations.
- Culture of “Swagger”: Finally Leveson points out that in some sectors safety is perceived to be an expression of weaknesses: ”Real men thrive on risk”.

Passed on this categorization provided by Leveson (N. Leveson, 2018), the impact of culture on safety is clearly delineated. These are pitfalls public administrations must be wary of. Managing safety culture is a continuous process that must be practiced and aligned throughout the organization. Not as a function of compliance but as an integrated aspect of design and operations. The lean expression of ”living safety” must be internalized in order for the previously outlined interventions to bear fruit. If practiced earnestly an assumption-based learning indicators program can help to enforce a culture of safety while monitoring drift in culture as a function of safety. With the active monitoring of assumptions, a shift towards a more safety-conscious mindset. This culture is characterized by readiness to adapt based on the feedback provided by the leading indicators. It encourages openness in discussing potential vulnerabilities and fosters a learning environment where safety becomes a shared responsibility. It shifts the organizational mindset to a continuous improvement process, a crucial change that will be discussed next.

The issue of the ”Fence”

Finally, we must address the general implications of applying system theory and system safety to the public administration domain. The issue is the separation of system development and operations. As observed from the Toeslagenaffair case, there seems to be a disconnect between system development and operations. We observed a system developed in a vacuum being deployed in a large batch into a very vast system landscape. As pointed out in chapter 1 this deployment results in ”adaptation” in the actual, real-world system landscape, resulting in the

emergence of new behavior. Through the disconnect between development and operations, this adaptation is not accounted for, and the integrity of assumptions made during the ”vacuum” design phase starts to fail. Inadvertently unsafe operational states are not detected and not rectified. This classical ”fence” approach between development and operations neglects an important fact true for any complex socio-technical system: *”Designers must assume the system to be dynamic and subject to continuous change”*. This observation brings the realization that the most important conclusion of this research must be, that future public ADM development and deployment in social welfare must follow a continuous project approach. Systems are never finished, they are just ”operational & safe” or not. Shifting to this long-term perspective during system development is crucial because it allows system developers to adopt a more sustainable approach to the lifecycle management of a system, they can introduce features in smaller batches and focus on improving the system over time based on feedback collected from operating the systems. This allows for system adaptation if design assumptions prove to be incorrect. Linking development and operations and removing barriers to their cooperation must be a focus of feature public administration management. This necessitates a long-term vision from decision-makers and the ability to break goals down into small deliverables, that are modular in nature.
To address these challenges, many organizations are adopting DevOps practices, a methodology that emphasizes collaboration and communication between software developers and other IT professionals while automating the process of software delivery and infrastructure changes. By integrating development and operations teams, social welfare agencies can achieve faster deployments and more reliable systems, thereby improving the administration of social services. In this thesis, we have proposed the addition of ”safety” to ensure development and operations are grounded in safe principles. Given the nature of decisions made in social welfare administration such an adaption is necessary to prevent technocracy from disproportionally scrutinizing marginalized groups of society (Alston, 2019). Implementing such practices requires significant cultural and procedural changes but can ultimately lead to more dynamic, responsive, efficient, and **safe** service delivery mechanisms within the social welfare domain.

## 8.2. Research Contributions

This thesis embarked on a journey to uncover the complexities and challenges associated with deploying Automated Decision-Making (ADM) systems within the social welfare administration. The inquiry was rooted in the realization that due to the complex characteristics of the social welfare system, the integration of ADM systems can inadvertently lead to the construction of digital cages. Thereby exacerbating challenges present in social welfare systems. This research aimed to demonstrate the utility of system safety thinking for the public administration domain. Through the analysis presented in this research, we have highlighted the systemic nature of digital cages and underscored the necessity for holistic, system safety approaches in the deployment and operation of ADM systems for social welfare.

Our findings concur with the argument that digital cages emerge not merely as a byproduct of technological implementations but as a manifestation of the systemic interplays within social welfare, see Peeters and Widlack (Widlak & Peeters, 2020). These cages, characterized by their rigid, exclusionary infrastructures, serve as barriers that prevent citizens from accessing the support and resources they rightfully deserve. Through our analysis of the issue from a system thinking perspective we have contributed to increasing the understanding of digital cages and the underlying causes for their emergence.

The thesis underscores the critical role of safety control interventions during the deployment and operation phases of ADM systems. It is during these stages that the potential for digital cages to emerge can be ’realized’, through the combination of environmental conditions and existing system hazards that are caused by inadequate control mechanisms. Thus, we argue this is where targeted interventions can have the most significant impact, on mitigating the emergence of digital cages. Our proposed solutions, grounded in system safety perspectives, aim to mitigate the risks associated with ADM system deployment and operation. By focusing on the control actions and interventions, we provide a framework for administrators to navigate the complexities of ADM systems in social welfare safely by utilizing assumption-based leading indicators.

Hereinafter we highlight the most relevant contributions made to the different research fields connected to this research.

### 8.2.1. Systems thinking for Public Administration

This thesis significantly enhances the understanding of ADM systems within the social welfare domain by advocating a holistic view that integrates insights from digital cage research. It underscores the necessity of considering the broader socio-technical environment in which these systems operate, highlighting how ADM systems interact with and influence broader social and system structures. By identifying emergent properties of complex systems that can lead to unintended consequences such as digital cages, the research contributes a nuanced perspective on the dynamics and interdependencies in a socio-technical context. This integration facilitates a deeper understanding of the systemic risks present in the social welfare domain that need to be considered during ADM system development and operations. The research provides a novel framework for analyzing complex systems, equipping policymakers and administrators with the tools to anticipate, identify, and mitigate systemic risks associated with ADM systems. This intervention design, topped by the assumption-based leading indicators, is particularly valuable for enhancing the robustness and reliability of administrative processes, ensuring that digital systems serve their intended purposes without triggering adverse societal impacts. The particular emphasis on system safety thinking allows for a methodology grounded in established theory, that shows promising results regarding its transferability to the social welfare domain.

### 8.2.2. Expansion of System Safety Research

This thesis expands the domain of system safety research by applying its principles to the digital and algorithmic contexts, particularly within the realm of social welfare systems. It introduces new dimensions of safety and risk, specific to digital infrastructures and algorithmic decision-making, through the integration of the digital cage concept into the system analysis. By developing targeted safety control interventions, utilizing STPA, the research delineates methods to identify and mitigate system hazards in large complex systems, thus preventing broader system failures and enhancing overall system safety. The methodologies and interventions developed in this research provide actionable guidance for the safe implementation of ADM systems. These insights are particularly crucial for complex and sensitive areas such as social welfare, where algorithmic decisions have profound impacts on human lives. By outlining specific safety interventions, the thesis aids in the design of ADM systems that are both effective and safe, safeguarding against potential misalignments between technological applications and social expectations. Furthermore, the designed interventions also highlight the potential of system safety methodology as an active problem-solving tool to overcome safety issues in administrative system development and operations.

### 8.2.3. Assumption-based leading indicators

This thesis has introduced assumption-based leading indicators as a systemic safety measure in the field of automated decision-making systems within social welfare systems. These indicators serve as early warning signals, alerting system administrators to potential breaches in safety constraints before they escalate into full-blown system failures or contribute to the unjust administrative exclusion of individuals from social welfare benefits. Through the strategic monitoring of these indicators, organizations can preemptively address emerging risks, thereby maintaining the integrity of the ADM and the overall system and ensuring it continues to operate within the boundaries of safety.

As outlined in this research assumption-based leading indicators can help to alleviate issues commonly encountered in operations of complex socio-technical systems and ADM systems. They provide operators with the ability to measure and assess the system state in uncertain and complex scenarios. They can help to circumvent the issue of causality, by enabling operators to gather information from the environment surrounding the model to make inferences regarding the 'safety' state of the model. More generally they are able to indicate the "migration to a higher level of risk" (N. Leveson, 2018). This research delineates a process of identifying, categorizing, and applying assumption-based leading indicators within social welfare administration. By doing so, not only enhances the operational safety of these systems but also contributes to a broader understanding of how ADM systems interact in the complex socio-technical system of social welfare. Furthermore, this thesis advocates for the integration of assumption-based leading indicators into the fabric of system design and management processes. It argues that such integration is essential for fostering a proactive safety culture within organizations, where continuous improvement and vigilance against potential system failures are ingrained values.

The theoretical and practical contributions of this work extend beyond the immediate context of social welfare systems. The principles and methodologies outlined here apply to a wide range of domains where ADM systems are deployed, offering valuable insights for researchers, practitioners, and policymakers interested in leveraging technology to serve public welfare while ensuring the system's safety and good governance.

The introduction of assumption-based leading indicators represents a significant step forward in ADM research, offering a novel tool for enhancing system reliability and safety. By setting forth assumption-based leading indicators, this thesis not only highlights the utility of system safety thinking but also pioneers a potential new research agenda in ADM. This agenda emphasizes the systematic identification, testing, and integration of underlying assumptions into the management of increasingly automated systems. The concept of assumption-based leading indicators broadens the scope of ADM research, urging a shift from reactive to proactive management practices fostering a deeper engagement with the complexities of socio-technical systems, and focusing on "safety boundaries" rather than product safety. It recognizes that safety is a product of a system in interaction with its environment, and thus system safety can only be achieved if the environmental conditions are accounted for.

In conclusion, the research on assumption-based leading indicators not only fills a gap in the existing literature on ADM system safety but also provides an impulse for future studies in this field. It underscores the importance of a nuanced understanding of design assumptions in safeguarding against unintended system failures and highlights the role of a robust safety control structure for the deployment and operations of technology within society.

## 8.3. Recommendations for Stakeholders

This work has shed light on the necessity for a shift in perspective, from viewing ADM systems as mere technological tools to recognizing them as integral components of a larger socio-technical system. This shift is essential for fostering an environment where technology serves to enhance, rather than hinder, the delivery of social welfare services. Based on the findings of this research several recommendations for different stakeholder groups can be derived. The following sections will highlight the relevant recommendations for each Stakeholder group.

### 8.3.1. Policy Makers

Policy Makers represent one of the most impactful stakeholders relevant to the issue of ADM in social welfare administration, as eluded to at many stages in this research, they have the power to shape the environmental conditions surrounding the system and provide direct input to the system through policy initiatives. Therefore they should incorporate the following recommendations:

Firstly, policymakers should integrate a holistic system safety approach in the design, deployment, and governance of ADM systems. This includes adopting frameworks that not only assess the technological aspects but also consider the socio-technical environment in which these systems operate. By understanding the interdependencies and potential emergent properties, policymakers can better anticipate and mitigate systemic risks.

Secondly, they should create regulatory frameworks that enforce transparency, accountability, and fairness in ADM applications, particularly those impacting social welfare. Regulations should require high standards regarding algorithmic causality tracing (version control, logging), data quality (source and ontology), and decision-making processes (business logic and mental model alignment). They should work towards audits and assessment frameworks that ensure compliance with these standards and societal values.

Finally, they should seek to find means to encourage collaboration between technologists, social scientists, safety, and legal experts in the policy-making process. This interdisciplinary approach can provide a comprehensive understanding of the implications of ADM systems and contribute to more robust policy frameworks that address multiple dimensions of system impact.

### 8.3.2. System Developers & Designers

System developers and designers are at the forefront of developing and deploying ADM systems into social welfare. The recommendations to them are vital to incorporate the essentials of safety thinking into future design work.

Developers should implement assumption-based leading indicators in the design and operational phases of ADM systems. These indicators will help monitor the system's adherence to expected behaviors and quickly identify deviations that could lead to failures or unintended consequences. Regular updates and reviews of these assumptions are critical as social contexts and technological landscapes evolve. Hence this program has to be introduced and managed in cooperation with system operators. This provides the segment to our second recommendation. Designers should shift toward a continuous deployment methodology that focuses on visualizing workflow, reducing batch size, and maintaining life-testing feedback environments. Such efforts should be strengthened with the implementation of ”DevSafOps” teams, that focus on integrating and aligning mental models and eliminating the ”fence” between system development and operations.

Coinciding with this they should increase the transparency of ADM systems by making system processes more understandable to non-technical stakeholders, including end users. We recommend designers implement an interdisciplinary problem-solving approach that adopts a holistic perspective, not merely focusing on the technical aspects of a system but rather the socio-technical implications it entails.

Finally, we urge the designers and public managers to review the current project management methodology. Specifically, we believe they should question the nature of contractual agreements with third-party developers and the incentive structure perpetuated through current ”business owner” and ”product owner” arrangements present in their organizations. These two areas are instrumental in aligning project goals early on yet are often rushed, ultimately resulting in a misalignment in an incentive structure that can result in harmful safety tradeoffs down the road. Keeping developers accountable and in-house product owners alignment is vital to any project's success.

### 8.3.3. Public Administrators & Operators

Public officials are also pivotal in the process of improving safety for ADM systems in social welfare because they are the system operators. Administrators should receive ongoing training on the latest developments in ADM and system safety methodology. This training will help them better understand the systems they are working with through adopting a ”system thinking” perspective. Moreover, system operators need to be vocal and advocate for necessary changes based on operational insights. Their assessment of the safety state of a system is fundamental to the effectiveness of the system safety approach.

To this purpose, they must set up comprehensive monitoring systems that utilize assumption-based leading indicators to continually assess the performance and impact of ADM systems and gauge the overall system state. These systems should be capable of triggering alerts when potential risks are identified, enabling timely interventions before issues escalate.

The corresponding alerts and general knowledge gathered from the operating process must be fed back into feedback loops with developers and policymakers to report back on system performance, societal impacts, and any challenges encountered in the operations of ADM systems. This feedback is crucial for iterative improvements and ensuring that systems remain aligned with their intended purposes.

Finally, we must street the implementation of contingency plans for the failure of systems in operations. As eluded to in this thesis, no system design is perfect, and if one thing is certain it is that nothing is certain. Hence, operators should have an actionable plan in place that allows the identification, interpretation, mitigation, and elimination of harm. Having a functional ”hedging process” in place significantly lowers the likelihood of harm arising out of a digital cage and should be a requirement before any system deployment.

### 8.3.4. Affected Citizens

For citizens potentially affected by the systems discussed in this research, we recommend maintaining a proactive attitude toward public service delivery. Citizens should informed and engaged in the public debate about these systems as much as possible. Staying vigilant regarding this subject will help scrutinize decision-makers while simultaneously improving one understanding of the rights and protections available to oneself.

Engaging in public discourse is a vital step. Participating in public consultations helps influence the implementation of these technologies by providing a citizen’s perspective on transparency and accountability. Citizens should advocate for transparency in the use of ADM systems by public agencies, demanding clarity on how decisions are made and how data is used. Supporting measures that hold developers and operators accountable and improve system safety are essential for ensuring these systems do not cause harm.

Finally, enhancing digital literacy is key. Understanding digital safety, and managing digital footprints can help mitigate risks associated with ADM systems. Opting out of unnecessary data collection and being cautious about online information sharing are practical steps to ensure personal data security and reduce the likelihood of misclassification algorithms. No matter if they are facilitated by government or privately owned systems. By taking these actions, citizens can contribute to shaping the development and deployment of ADM technologies in a way that respects democratic values and human rights, ultimately helping former Special Rapporteur Philip Alston to sleep a bit easier.

### 8.3.5. Academia

Academia is another important stakeholder that through indirect influence can help shape the long-term evolution of ADM in social welfare. Researchers should engage in interdisciplinary research to explore the complexities and challenges at the intersection of technology, society, and policy. They should aim to provide insights that inform better design practices, enhance system safety methodologies, and contribute to effective policy-making.

Secondly, they should carry on the work of developing and refining methodologies that can assess and enhance the safety and efficacy of ADM systems. This includes expanding the use of assumption-based leading indicators and exploring new methods for predicting and mitigating emergent risks.

Finally, they should advocate for open science practices and collaborative research initiatives with public administrators. During this research, we experienced the segmentation present between scientific research and public administration for this sensible issue. Researchers in a position of influence should help facilitate the sharing of findings, tools, and methodologies across the sectors. This collaboration can accelerate innovation and the adoption of best practices in ADM system development and deployment for social welfare.

## 8.4. Future Directions & Research

In conclusion, this thesis not only highlights the challenges posed by the integration of ADM systems in social welfare but also offers a path forward. By adopting a system safety perspective and prioritizing the deployment and operation phases, we can pave the way for a future where technology empowers rather than entraps those it is meant to serve. To this end, we once more want to highlight the need for interdisciplinary research that helps deepen our understanding of this complex subject.

Moving forward, it is imperative that research and practice in the deployment and operation of ADM systems in social welfare adopt a more nuanced, systems-oriented approach. This thesis contributes to the groundwork for further exploration into the design and implementation of safety control interventions. Future research should delve into the implications of this thesis. Hereinafter we will delineate specific research directions that are worth exploring.

### 8.4.1. System Safety Theory for Public Administration

Utility of System Safety for Public Administration

Additional effort should be directed towards proving the utility of system safety thinking to modern public administration that ensures alignment to the particularities of this domain and, if necessary, adapts current system safety methodology to accommodate these. Several of these aspects have been highlighted in the lessons learned section on the application of STPA in public administration. The additional effort could be commenced by applying STPA in the form of real-world case studies that can provide ex ante and ex post evaluation results. This quantitative information would significantly improve the likelihood of convincing relevant stakeholders. Particular emphasis should be placed on proving the potential for ‘low cost’ safety improvements show showcasing the ability of system safety methodology to generate ‘quick wins’ that do not have to come at the cost of a scandal, loss of reputation, or increased scrutiny on those responsible. As mentioned by Leveson, system safety can help shift culture away from the ”blame game” (N. Leveson, 2018).

Utility of control interventions

Future research should delve into the practical application of the interventions proposed in this research, exploring their effectiveness in real-world settings and identifying areas for refinement. Such research is pivotal to increasing the system designer’s practical ability to combat safety issues. In this regard, researching and proving the utility of the proposed “hedging process” should be prioritized. Such a standardized process can harness great mitigation potential, vital to ‘braking’ ripple effects and thereby hindering the emergence of digital cages.

Integrating System Thinking in Public Administration

Finally, future research must address the specific way in which the system think perspective, underlying system safety methodology, can be integrated into modern public administration. The concepts conveyed through this perspective are complex and hence require a comprehensive, directed training approach. To ‘place’ this training correctly, a socio-technical analysis of the government structure could be conducted with the goal of identifying specific ‘key stakeholders’ able to facilitate this perspective into different administrative project initiatives after they have received training.

### 8.4.2. Assumption-based leading indicators for ADMs

One fundamental contribution of this research is the introduction of assumption-based leading indicators as a system safety tool for automated decision-making systems. Subsequently, we will highlight future areas of research that result from this finding.

Leading Indicators for Public Administration

The introduction of an assumption-based leading indicator program for public ADM systems in this research represents a conceptual demonstration. However, it does not provide quantitative data regarding its utility. This approach has to be evaluated further with specific assumption-based leading indicators. Due to the depth of this topic, these have been kept rather arbitrary in this thesis. Future research should focus on investigating potential, specific further leading indicators and prove their utility. Such research could be conducted by identifying specific indicators such as system “workload” and conducting simulations such as discrete event simulations to test how a specific system behaves under different loads. Note that with ‘system’ we refer to the socio-technical system/process tasked with delivering a specific objective, not the isolated technical ADM system, as such a test would not capture the needed sub-system interactions.

Leading Indicators for other Domains

Finally, as demonstrated by Leveson in her research on Assumption-based leading indicators, they are applicable to a wide range of domains. Yet identifying general indicators applicable to a wide number of problems present in specific domains has in the past proven to be unfruitful (N. Leveson, 2015). Therefore, given the inherent potential of assumption-based leading indicators to improve ADM safety, research into subject-specific indicators for other areas of application for ADM systems should be commenced. Such research may be generalized to domains utilizing AI technology. Given for example the recent rise of large language models research into leading indicators that are able to project model drift could prove useful. Given the assumption that AI applications are in a state of constant change in combination with their “black box” characteristics (problem of establishing causality), assumption-based leading indicators could be pivotal to AI safety, because the assumption-based leading indicator program does not have to measure a holistic model state. It rather operates on the assumption that if a specific model indicator is “triggered” the state of the system can assumed to be fragile, necessitating intervention. Identifying such indicators, however, will necessitate closer cooperation between current system developers and researchers in the field of system safety.

# 9

# Final remarks

To conclude this Thesis some final remarks:

While it can be concluded that the safe deployment principles, despite their limitations, can indeed improve the safety of the social welfare system by inhibiting the emergence of digital cages, they are by no means sufficient in and of themselves. A central issue at the heart of this problem remains almost untouched by them. Previously we have discussed the larger macro trends leading up to the increased utilization of ADM systems in social welfare. The often cited and most significant factor for the adoption of ADM systems by governments is the desire to create “effectiveness” and “efficiency” gains. Crucially, however, in reality, this desire is often quantified through a one-dimensional lens, “profitability”. The “effectiveness” and “efficiency” of a project or policy are assessed by quantifying its potential for monetary gains or savings and thereby inadvertently transformed into a “business case”. Such a frame of analysis however fails to capture essential values inherent to any functional social welfare system. It fails to acknowledge that the purpose of any welfare policy or system is to serve people by providing them with welfare and social security. The matter of digital cages is therefore not only a matter of improving socio-technical processes within the social welfare system. It is also a matter of questioning the purpose of institutions, policy, and the thereout resulting projects. Is a system designed to regulate, punish, and discipline or is it designed to support, secure and strengthen? In the day and age of artificial intelligence, big data, and “rationalization” the curse of flexibility is bearing down on us. Consequently, the question policymakers, designers, and developers should ask is not: “How can we design this system?” but rather “Should this system exist?”.

# Bibliography

Alkhatib, A. (2021). To Live in Their Utopia: Why Algorithmic Systems Create Absurd Outcomes. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–9. https://doi.org/10.1145/3411764.3445740

Allen, T. F. H., Tainter, J. A., & Hoekstra, T. W. (1999). Supply-Side Sustainability. *Systems Research and Behavioral Science*, (16), 403–427.

Alston, P. (2019, November). *Extreme poverty and human rights : Note / by the Secretary-General* (tech. rep. No. A/74/493). United Nations. New York, NY, USA. https://digitallibrary.un.org/record/3834146?ln=en

Ashby, R. W. (n.d.). Instructie FSV. Archivering documenten Toezichtlijst en Beheer Uitsluitingen TVS. Retrieved January 31, 2024, from https://www.tweedekamer.nl/kamerstukken/detail

Assembly, U. N. G. (2015, October). 70/1. Transforming our world: The 2030 Agenda for Sustainable Development. Retrieved September 5, 2023, from https://documents-dds-ny.un.org/doc/UNDOC/GEN/N15/291/89/PDF/N1529189.pdf?OpenElement

Axelrod, R., & Hamilton, W. D. (1981). The Evolution of Cooperation [Publisher: American Association for the Advancement of Science]. *Science*, *211*(4489), 1390–1396. https://doi.org/10.1126/science.7466396

Baehr, P. (2001). The “Iron Cage” and the “Shell as Hard as Steel”: Parsons, Weber, and the Stahlhartes Gehäuse Metaphor in the Protestant Ethic and the Spirit of Capitalism [_eprint: https://onlinelibrary.wiley.com/doi/ 2656.00160]. *History and Theory*, *40*(2), 153–169. https://doi.org/10.1111/0018-2656.00160

Bannister, F., & Connolly, R. (2020). Administration by algorithm: A risk management framework [Publisher: IOS Press]. *Information Polity*, *25*(4), 471–490. https://doi.org/10.3233/IP-200249

Barth, T. J., & Arnold, E. (1999). Artificial Intelligence and Administrative Discretion: Implications for Public Administration [Publisher: SAGE Publications Inc]. *The American Review of Public Administration*, *29*(4), 332–351. https://doi.org/10.1177/02750749922064463

Belastingdienst. (n.d.). Data & Analyse. Retrieved April 5, 2024, from https://werken.belastingdienst.nl/data-en-analyse-oud

Belastingdienst. (2022, June). Missie en visie. Retrieved April 5, 2024, from https://over-ons.belastingdienst.nl/organisatie/missie-visie/

Bovens, M., & Zouridis, S. (2002). From Street-Level to System-Level Bureaucracies: How Information and Communication Technology is Transforming Administrative Discretion and Constitutional Control [_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/0033-3352.00168]. *Public Administration Review*, *62*(2), 174–184. https://doi.org/10.1111/0033-3352.00168

Bozeman, B. (1993, July). A Theory of Government ”Red Tape” on JSTOR. Retrieved April 27, 2023, from https://www.jstor.org/stable/1181785

Bozeman, B. (2000). *Bureaucracy and red tape* (Vol. 14). Prentice Hall Upper Saddle River, NJ.

Brazier, F., van Langen, P., Lukosch, S., & Vingerhoeds, R. (2018). Complex Systems: Design, engineering, governance. In H. L. Bakker & J. P. de Kleijn (Eds.), *Projects and People* (pp. 35–60). NAP. Retrieved September 5, 2023, from https://oatao.univ-toulouse.fr/23817/

Breck, E., Cai, S., Nielsen, E., Salib, M., & Sculley, D. (2017). The ML test score: A rubric for ML production readiness and technical debt reduction. *2017 IEEE International Conference on Big Data (Big Data)*, 1123–1132. https://doi.org/10.1109/BigData.2017.8258038

Brodkin, E. Z., & Majmundar, M. (2010). Administrative Exclusion: Organizations and the Hidden Costs of Welfare Claiming. *Journal of Public Administration Research and Theory*, *20*(4), 827–848. https://doi.org/10.1093/jopart/mup046

Bullock, J. B. (2019). Artificial Intelligence, Discretion, and Bureaucracy [Publisher: SAGE Publications Inc]. *The American Review of Public Administration*, *49*(7), 751–761. https://doi.org/10.1177/0275074019856123

Bungay, S. (2021, October). The Art of Action bei hugendubel.de. Online bestellen oder in der Filiale abholen. [ISBN: 9781529376968]. Retrieved July 11, 2023, from https://www.hugendubel.de/de/buch_gebunden/stephen_bungay-the_art_of_action-41326133-produkt-details.html

Capgemini. (2013, January). Ontwikkelspecificatie Dagboek Fraude Signalering Voorziening FSV SOS NET Dienstverlening Belastingdienst versie 1.1.

Collington, R. (2022). Disrupting the Welfare State? Digitalisation and the Retrenchment of Public Sector Capacity [Publisher: Routledge _eprint: https://doi.org/10.1080/13563467.2021.1952559]. *New Political Economy*, *27*(2), 312–328. https://doi.org/10.1080/13563467.2021.1952559

Corning, P. (2015, May). The Re-emergence of “Emergence”: A Venerable Concept in Search of a Theory | Institute for the Study of Complex Systems. Retrieved September 24, 2023, from https://complexsystems.org/publications/the-re-emergence-of-emergence-a-venerable-concept-in-search-of-a-theory/

Côté, P.-O., Nikanjam, A., Bouchoucha, R., Basta, I., Abidi, M., & Khomh, F. (2023, June). Quality Issues in Machine Learning Software Systems [arXiv:2306.15007 [cs]]. Retrieved February 7, 2024, from http://arxiv.org/abs/2306.15007

Dall-E. (2024, May). ChatGPT - image generator. Retrieved April 5, 2024, from https://chat.openai.com/g/g-pmuQfob8d-image-generator

Djeffal, C. (2020). Artificial Intelligence and Public Governance: Normative Guidelines for Artificial Intelligence in Government and Public Administration. In T. Wischmeyer & T. Rademacher (Eds.), *Regulating Artificial Intelligence* (pp. 277–293). Springer International Publishing. https://doi.org/10.1007/978-3-030-32361-5_12

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

Dobbe, R. I. J. (2022). System Safety and Artifcial Intelligence □, 15. https://doi.org/https://doi-org.tudelft.idm.oclc.org/10.1093/oxfordhb/9780197579329.001.0001

Edwards, L., & Veale, M. (2017). Slave to the Algorithm? Why a ’Right to an Explanation’ Is Probably Not the Remedy You Are Looking For. *Duke Law & Technology Review*, *16*(1), 18–84. https://scholarship.law.duke.edu/dltr/vol16/iss1/2

Etscheid, J. (2019). Artificial Intelligence in Public Administration. In I. Lindgren, M. Janssen, H. Lee, A. Polini, M. P. Rodríguez Bolívar, H. J. Scholl, & E. Tambouris (Eds.), *Electronic Government* (pp. 248–261). Springer International Publishing. https://doi.org/10.1007/978-3-030-27325-5_19

Filgueiras, F. (2022). New Pythias of public administration: Ambiguity and choice in AI systems as challenges for governance. *AI & SOCIETY*, *37*(4), 1473–1486. https://doi.org/10.1007/s00146-021-01201-4

Financiën, M. v. (2022, January). Onderzoek Gegevensdeling met Derden - Rapport - Rijksoverheid.nl [Last Modified: 2022-03-23T14:03 Publisher: Ministerie van Algemene Zaken]. Retrieved January 30, 2024, from https://www.rijksoverheid.nl/documenten/rapporten/2022/01/19/bijlage-2-rapport-pwc-externe-gegevensdeling-uit-fsv

Floridi, L., & Cowls, J. (2019). A Unified Framework of Five Principles for AI in Society. *Harvard Data Science Review*. https://doi.org/10.1162/99608f92.8cd550d1

Gordon, T. F. (2004). eGovernance and its Value for Public Administration. *KDNet Symposium on Knowledge-Based Services for the Public Sector: Bonn, June*, 3–4.

Graycar, A., & Masters, A. B. (2022). Bureaucratic bastardry: Robodebt/debt recovery, AI and the stigmatisation of citizens by machines and systems [Publisher: Inderscience Publishers]. *International Journal of Public Policy*, *16*(5-6), 333–344. https://doi.org/10.1504/IJPP.2022.127432

Hayes, B. (2015, November). Ben Hayes - Demystifying the Confusion Matrix. Retrieved April 12, 2024, from https://benhay.es/posts/demystifying-confusion-matrix/

Heikkilä, M. (2022, March). Dutch scandal serves as a warning for Europe over risks of using algorithms. Retrieved September 20, 2023, from https://www.politico.eu/article/dutch-scandal-serves-as-a-warning-for-europe-over-risks-of-using-algorithms/

Henriques-Gomes, L. (2023). Robodebt: Five years of lies, mistakes and failures that caused a $1.8bn scandal. *The Guardian*. Retrieved April 27, 2023, from https://www.theguardian.com/australia-news/2023/mar/11/robodebt-five-years-of-lies-mistakes-and-failures-that-caused-a-18bn-scandal

Hevner, A. R., March, S. T., Park, J., & Ram, S. (2004). Design Science in Information Systems Research.

Jakubiec, M. (2022). Legal Concepts as Mental Representations. *International Journal for the Semiotics of Law - Revue internationale de Sémiotique juridique*, *35*(5), 1837–1855. https://doi.org/10.1007/s11196-021-09853-7

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

Johannesson, P., & Perjons, E. (2014, September). *An Introduction to Design Science*. Springer. Retrieved April 6, 2024, from https://link.springer.com/book/10.1007/978-3-319-10632-8

Jonk, E., & Iren, D. (2021). Governance and Communication of Algorithmic Decision Making: A Case Study on Public Sector. *2021 IEEE 23rd Conference on Business Informatics (CBI)*, 151–160. https://doi.org/10.1109/CBI52690.2021.00026

Jorna, F., & Wagenaar, P. (2007). The 'Iron Cage' Strengthened? Discretion and Digital Discipline [_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1467-9299.2007.00640.x]. *Public Administration*, *85*(1), 189–214. https://doi.org/10.1111/j.1467-9299.2007.00640.x

Karkliniewska, I. (2022). Building Transparency and Robustness of AI/ADM Management in Public Sector.

Kim, G., Debois, P., Willis, J., & Humble, J. (2016a). AgileJazz - Russ Wangler's Blog: DevOps Handbook Summary 2 of 4 - The Second Way. Retrieved March 27, 2024, from https://agilejazz.blogspot.com/p/the-devops-handbook-summary-2-of-4.html

Kim, G., Debois, P., Willis, J., & Humble, J. (2016b, September). *The DevOps Handbook: How to Create World-Class Agility, Reliability, and Security in Technology Organizations*. IT Revolution Press.

KPMG. (2020). Rapportage verwerking van risicosignalen voor toezicht. *10*.

Lamar, K., Sounwave, Thundercat, & Kuti, F. (2015, March). Mortal Man, To Pimp a Butterfly, Format: BANGER! https://genius.com/Kendrick-lamar-mortal-man-lyrics

Le Dantec, C. A., & Edwards, W. K. (2010). Across boundaries of influence and accountability: The multiple scales of public sector information systems. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 113–122. https://doi.org/10.1145/1753326.1753345

Leveson, N. (2011). *Engineering a safer world: Systems thinking applied to safety* [OCLC: ocn719429220]. MIT Press.

Leveson, N. (2015). A systems approach to risk management through leading safety indicators. *Reliability Engineering & System Safety*, *136*, 17–34. https://doi.org/10.1016/j.ress.2014.10.008

Leveson, N. (2018). STPA Handbook.

Leveson, N. G. (2019). CAST Handbook.

Lin, S.-H., & Ikram, M. A. (2020). On the relationship of machine learning with causal inference. *European Journal of Epidemiology*, *35*(2), 183–185. https://doi.org/10.1007/s10654-019-00564-9

Loi, M., & Spielkamp, M. (2021). Towards Accountability in the Use of Artificial Intelligence for Public Administrations [arXiv:2105.01434 [cs]]. *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, 757–766. https://doi.org/10.1145/3461702.3462631

Lorenz, L., Meijer, A., & Schuppan, T. (2021). The algocracy as a new ideal type for government organizations: Predictive policing in Berlin as an empirical case [Publisher: IOS Press]. *Information Polity*, *26*(1), 71–86. https://doi.org/10.3233/IP-200279

Mao, F. (2023). Robodebt: Illegal Australian welfare hunt drove people to despair. *BBC News*. Retrieved September 20, 2023, from https://www.bbc.com/news/world-australia-66130105

Miro. (n.d.). Ideen Parkplatz-Matrix Vorlage & Beispiel für Teams | Miro. Retrieved April 5, 2024, from https://miro.com/de/templates/ideen-parkplatzmatrix/

Moran, M., Rein, M., & Goodin, R. E. (Eds.). (2006). *The Oxford handbook of public policy* [OCLC: ocm62878681]. Oxford University Press.

Mulligan, D. K., & Bamberger, K. A. (2019). Procurement As Policy: Administrative Process for Machine Learning. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3464203

Myrbakken, H., & Colomo-Palacios, R. (2017). DevSecOps: A Multivocal Literature Review. In A. Mas, A. Mesquida, R. V. O'Connor, T. Rout, & A. Dorling (Eds.), *Software Process Improvement and Capability Determination* (pp. 17–29). Springer International Publishing. https://doi.org/10.1007/978-3-319-67383-7_2

Nouws, S., Janssen, M., & Dobbe, R. (2022). Dismantling Digital Cages: Examining Design Practices for Public Algorithmic Systems. In M. Janssen, C. Csáki, I. Lindgren, E. Loukis, U. Melin, G. Viale Pereira, M. P. Rodríguez Bolívar, & E. Tambouris (Eds.), *Electronic Government* (pp. 307–

322). Springer International Publishing. https://link-springer-com.tudelft.idm.oclc.org/chapter/10.1007/978-3-031-15086-9_20

Ojo, A., Mellouli, S., & Ahmadi Zeleti, F. (2019). A Realist Perspective on AI-era Public Management*. *Proceedings of the 20th Annual International Conference on Digital Government Research*, 159–170. https://doi.org/10.1145/3325112.3325261

Oosthuizen, R. M. (2022). The Fourth Industrial Revolution – Smart Technology, Artificial Intelligence, Robotics and Algorithms: Industrial Psychologists in Future Workplaces. *Frontiers in Artificial Intelligence*, *5*. Retrieved July 14, 2023, from https://www.frontiersin.org/articles/10.3389/frai.2022.913168

Pandey, S. K., & Scott, P. G. (2002). Red Tape: A Review and Assessment of Concepts and Measures. *Journal of Public Administration Research and Theory*, *12*(4), 553–580. https://doi.org/10.1093/oxfordjournals.jpart.a003547

Peeters, R., & Widlak, A. (2018). The digital cage: Administrative exclusion through information architecture – The case of the Dutch civil registry's master data management system. *Government Information Quarterly*, *35*(2), 175–183. https://doi.org/10.1016/j.giq.2018.02.003

Peeters, R., & Widlak, A. C. (2023). Administrative exclusion in the infrastructure-level bureaucracy: The case of the Dutch daycare benefit scandal [_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/puar.1361 *Public Administration Review*, *83*(4), 863–877. https://doi.org/10.1111/puar.13615

Peffers, K., Tuunanen, T., Rothenberger, M. A., & Chatterjee, S. (2007). A Design Science Research Methodology for Information Systems Research [Publisher: Routledge _eprint: https://doi.org/10.2753/MIS0742-1222240302]. *Journal of Management Information Systems*, *24*(3), 45–77. https://doi.org/10.2753/MIS0742-1222240302

Pel, L. (2022). Ripple Effects of Law Execution Automation in Governmental Systems.

Persoonsgegevens, A. (2021, October). *Verwerkingen van persoonsgegevens in de Fraude Signalering Voorziening (FSV)* (tech. rep.). https://www.autoriteitpersoonsgegevens.nl/documenten/onderzoek-belastingdienst-fraude-signalering-voorziening-fsv

PwC. (2021, November). Onderzoek effecten FSV Toeslagen - Rapport - Rijksoverheid.nl [Last Modified: 2022-03-14T11:12 Publisher: Ministerie van Algemene Zaken]. Retrieved January 29, 2024, from https://www.rijksoverheid.nl/documenten/rapporten/2021/12/03/onderzoek-pwc-effecten-fsv-toeslagen

Ranerup, A., & Svensson, L. (2023). Automated decision-making, discretion and public values: A case study of two municipalities and their case management of social assistance [Publisher: Routledge _eprint: https://doi.org/10.1080/13691457.2023.2185875]. *European Journal of Social Work*, *0*(0), 1–15. https://doi.org/10.1080/13691457.2023.2185875

Rasmussen, J. (1997). Risk management in a dynamic society: A modelling problem [ISBN: 0925-7535 Publisher: Elsevier]. *Safety science*, *27*(2-3), 183–213.

Sculley, D., Holt, G., Golovin, D., Davydov, E., Phillips, T., Ebner, D., Chaudhary, V., Young, M., Crespo, J.-F., & Dennison, D. (2015). Hidden Technical Debt in Machine Learning Systems. *Advances in Neural Information Processing Systems*, *28*. Retrieved February 7, 2024, from https://proceedings.neurips.cc/paper_files/paper/2015/hash/86df7dcfd896fcaf2674f757a2463eba-Abstract.html

Spielkamp, M. (2019). Automating Society: Taking Stock of Automated Decision-Making in the EU. BertelsmannStiftung Studies 2019.

Suresh, H., & Guttag, J. V. (2021). A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle [arXiv:1901.10002 [cs, stat]]. *Equity and Access in Algorithms, Mechanisms, and Optimization*, 1–9. https://doi.org/10.1145/3465416.3483305

Thoring, K., Mueller, R., & Badke-Schaub, P. (2020). Workshops as a Research Method: Guidelines for Designing and Evaluating Artifacts Through Workshops. https://doi.org/10.24251/HICSS.2020.620

Times, N. (2021, October). Over 1,100 children taken from homes of benefits scandal victims | NL Times. Retrieved September 20, 2023, from https://nltimes.nl/2021/10/19/1100-children-taken-homes-benefits-scandal-victims

Toeslagenaffaire [Page Version ID: 67200002]. (2024, March). Retrieved March 23, 2024, from https://nl.wikipedia.org/w/index.php?title=Toeslagenaffaire&oldid=67200002

van Engers, T. M., & de Vries, D. M. (2019). Governmental Transparency in the Era of Artificial Intelligence. *JURIX*, 33–42. https://ebooks.iospress.nl/volumearticle/53651 van Noordt, C., & Misuraca, G. (2020). Evaluating the impact of artificial intelligence technologies in public services: Towards an assessment framework. *Proceedings of the 13th International Conference on Theory and Practice of Electronic Governance*, 8–16. https://doi.org/10.1145/3428502.3428504

Veale, M., & Brass, I. (2019). Administration by Algorithm?

Veale, M., Van Kleek, M., & Binns, R. (2018). Fairness and Accountability Design Needs for Algorithmic Support in High-Stakes Public Sector Decision-Making. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 1–14. https://doi.org/10.1145/3173574.3174014

Waldrop, M. M. (1993). *Complexity: The emerging science at the edge of order and chaos*. Simon; Schuster.

Weber, M. (2016). *Die protestantische Ethik und der ”Geist” des Kapitalismus* (K. Lichtblau & J. Weiß, Eds.). Springer Fachmedien. https://doi.org/10.1007/978-3-658-07432-6

Weber, M., & Tawney, R. H. (1930). *The Protestant ethic and the spirit of capitalism* (T. Parsons, Trans.) [OCLC: 3064861]. George Allen & Unwin Ltd., Museum Street.

Weyerer, J. C., & Langer, P. F. (2019). Garbage In, Garbage Out: The Vicious Cycle of AI-Based Discrimination in the Public Sector. *Proceedings of the 20th Annual International Conference on Digital Government Research*, 509–511. https://doi.org/10.1145/3325112.3328220

Widlak, A., & Peeters, R. (2020). Administrative errors and the burden of correction and consequence: How information technology exacerbates the consequences of bureaucratic mistakes for citizens [Publisher: Inderscience Publishers]. *International Journal of Electronic Governance*, *12*(1), 40–56. https://doi.org/10.1504/IJEG.2020.106998

Williamson, O. E. (1998). Transaction cost economics: How it works; where it is headed [ISBN: 0013-063X Publisher: Springer]. *De economist*, *146*, 23–58.

Wirtz, B. W., Langer, P. F., & Fenner, C. (2021). Artificial Intelligence in the Public Sector - a Research Agenda [Publisher: Routledge _eprint: https://doi.org/10.1080/01900692.2021.1947319]. *International Journal of Public Administration*, *44*(13), 1103–1128. https://doi.org/10.1080/01900692.2021.1947319

Wirtz, B. W., & Müller, W. M. (2019). An integrated artificial intelligence framework for public management [Publisher: Routledge _eprint: https://doi.org/10.1080/14719037.2018.1549268]. *Public Management Review*, *21*(7), 1076–1100. https://doi.org/10.1080/14719037.2018.1549268

Wirtz, B. W., Weyerer, J. C., & Sturm, B. J. (2020). The Dark Sides of Artificial Intelligence: An Integrated AI Governance Framework for Public Administration [Publisher: Routledge _eprint: https://doi.org/10.1080/01900692.2020.1749851]. *International Journal of Public Administration*, *43*(9), 818–829. https://doi.org/10.1080/01900692.2020.1749851

Young, M. M., Bullock, J. B., & Lecy, J. D. (2019). Artificial Discretion as a Tool of Governance: A Framework for Understanding the Impact of Artificial Intelligence on Public Administration. *Perspectives on Public Management and Governance*, *2*(4), 301–313. https://doi.org/10.1093/ppmgov/gvz014

Young, M. M., Himmelreich, J., Bullock, J. B., & Kim, K.-C. (2021). Artificial Intelligence and Administrative Evil. *Perspectives on Public Management and Governance*, *4*(3), 244–258. https://doi.org/10.1093/ppmgov/gvab006

# A

# Results from literature review

| Article | Perpetuating Socio-technial Factors | Essential Design Advise |
|---|---|---|
| (Alkhatib, 2021) | **- Uneven power structures in the design of algorithmic**<br>**- Historical biases in data reinforce existing power structures**<br>**- Post-processing biases through misuse of ADM tool**<br>**- Lack of transparency and accountability** | - ”disempower” ADM systems<br>- use participatory design methodology |
| (Bannister & Connolly, 2020) | **- Lack of democratic participation in design**<br>**- Opacity of ADM systems makes challenging/understanding decisions difficult**<br>**- Lack of transparency and accountability**<br>**- Power of none-government system designers (privatization)**<br>**- Lack of education on the effects and risks of ADM system in users and managers** | - Improve detection and categorization capabilites of risks surrounding ADM systems<br>- Redesign incentive structure in ADM design |
| (Barth & Arnold, 1999) | **- Increased complexity of government operations**<br>**- ADM as black boxes to citizens and bureaucrats leading to a lack of oversight and control over decision making**<br>**- ADM systems obstruct discretion**<br>**- Power of none-government system designers (privatization)** | - Stronger contexualized AI knowledge in public administration |
| (Bullock, 2019) | **- Obstruction of administrative discreation to make circumstance sensitive decision**<br>**- Lack of transparency and accountability**<br>**- Power of none-government system designers (privatization)** | - Better unstanding what tasks are viable for ADM automation<br>- Participatory and value oriantated design |
| (Collington, 2022) | **- Administrative exclusion of citizens not able to access digital technologies**<br>**- Power of none-government system designers (privatization)**<br>**- The erosion of public sector capacity** | - Changing the incentive strucuture for public contractors<br>- Reversing retrenchment |
| (Djeffal, 2020) | **- Lack of transparency and accountability**<br>**- Historical and other biases in data**<br>**- Alienation and disempowerment of Users though lack of participation rights**<br>**- Uneven power structures in the design of algorithmic** | - Human-centered approach to the governance of AI in public administration<br>- Strucuturing a processes to arrive at normative guidelines |
| (R. I. J. Dobbe, 2022) | **- Increasing complexity of AI/ADM systems hinders in establishing causality**<br>**- Lack of transparency and accountability**<br>**- Alienation and disempowerment of Users though lack of participation rights**<br>**- Lack of system safety perspective in ADM design**<br>**- Historical and other biases in data**<br>**- Deployment of ADM without imposed safety contraints** | - Mindfulness of ADM system constraints<br>- Alignment mental models\<br>- Build a culture open to understanding and learning<br>- Adopting design loops |

| Article | Perpetuating Socio-technial Factors | Essential Design Advise |
|---|---|---|
| (Edwards & Veale, 2017) | **- Historical and other biases in data**<br>**- Flawed design practices**<br>**- Lack of transparency and accountability** | - XAI but not without:<br>- Participatory and adaptive design<br>- Privacy by design<br>- Data Protection Impact Assessments<br>- Certification and Privacy seals |
| (Etscheid, 2019) | **- Lack of democratic participation and diversity in design**<br>**- Lack of public citizens education about ADM systems**<br>**- Power of none-government system designers (privatization)** | - user centric AI design<br>- reconfigurating insentive structure |
| (Filgueiras, 2022) | **- Lack of transparency and accountability**<br>**- Histocial and other biases in data**<br>**- Lack of user participation in design**<br>**- Insufficant normative guidelines for ADM design** | - Design institutions that can establish norms and policies for democratically guiding actors in designing and adopting ADM systems |
| (Graycar & Masters, 2022) | **- Lack of participatory design methodology**<br>**- Lack of transparency and accountability**<br>**- Opacity of ADM systems makes challenging/understanding decisions difficult** | - transparency and public scrutiny<br>- integration of anti-coruption frameworks |
| (Janssen & Kuk, 2016) | **- Technocratic systems design approach**<br>**- Flawed data curation methods**<br>**- Lack of transparency and accountability** | - democratize data curatorial practices through greater openness and transparency in data procurement, custody, and utilization. |
| (Jonk & Iren, 2021) | **- Lack of risk aware design methdologies**<br>**- Lack of transparency and accountability**<br>**- Lack of democratic participation and diversity in design**<br>**- Missing normative guidelines and regulations for ADM systems** | - clearer, flexible and inclusive methodological framework |
| (Jorna & Wagenaar, 2007) | **- Increased datafication of public administration**<br>**- Increased complexity of decision making**<br>**- Technocratic system design approach**<br>**- Lack of transparency and accountability** | - Design can not solely focus on artefacts<br>- Have entire system and its participatory boundary practices in mind |
| (Karkliniewska, 2022) | **- Lack of transparency and accountability**<br>**- Lack of participatory design methodology**<br>**- Lack of user control/influence in deployment**<br>**- Lack of democratic participation and diversity in design** | - Ongoing monitoring and evaluation of AI/ADM systems<br>- Ensure effectiveness and ethical use |
| (Le Dantec & Edwards, 2010) | **- Data & Information silos obstruct causality and impact tracking**<br>**- Lack of transparency and accountability**<br>**- Power of none-government system designers (privatization)** | - Calls for more research into multi-level and segmented ICT systems<br>- Improvement of design process to adopt to organisational needs on different levels of aggregation |
| (Loi & Spielkamp, 2021) | **- Data & Information silos obstruct causality and impact tracking**<br>**- Lack of transparency and accountability**<br>**- Historical and other biases in data**<br>**- Lack of user control and intervention mechanisms** | - Accountability and Auditability<br>- Greater control over ADM decision outcomes and processes |

| Article | Perpetuating Socio-technial Factors | Essential Design Advise |
|---|---|---|
| (Lorenz et al., 2021) | **- The increasing use of algorithmic decision-making**<br>**- Increased centralization of decision-making**<br>**- Lack of transparency and accountability**<br>**- Uneven power structures in the design of algorithmic** | - Algocracy as an ideal type overgovernance<br>- Ensure centralization does not effect discretion |
| (Mulligan & Bamberger, 2019) | **- Increased complexity of machine learning/ ADM algorithms**<br>**- Lack of democratic participation and diversity in design**<br>**- Lack of transparency and accountability**<br>**- Historical and other biases in data**<br>**- Power of none-government system designers (privatization)** | - Move from a procurement mindset to policymaking mindset<br>- Participatory, iterative design to allow for ”political visibility” and “contestable design”<br>- Restructuring of incentive structure surrounding ADM design<br>- Improved training of public officials in the development, management and use of ADM systems |
| (Nouws et al., 2022) | **- Narrow focus on technical artefacts**<br>**- Disregard for normative basis for ADM systems**<br>**- Depend on involved actors’ awareness of socio-technical components and interactions in public algorithmic systems**<br>**- Linear rather than iterative** | - Design ADM in its sociotechnical context<br>- Set clear normative framework for ADM use<br>- Employ inclusive design methodology<br>- Move from waterfall to iterative design process |
| (Ojo et al., 2019) | **- Technological determinism**<br>**- Lack of transparency and accountability**<br>**- Power of none-government system designers (privatization)** | - Design ADM in its sociotechnical context<br>- Move away form a NPM organisation focused design |
| (Peeters & Widlak, 2018) | **- ADM as black boxes to citizens and bureaucrats leading to a lack of oversight and control over decision making**<br>**- Legal contamination via ICT**<br>**- ICT eliminating street level discretion**<br>**- Perverse behavioural incentives for administrative bodies** | - Focus on detecting and eliminating ’ripple effects’<br>- Eliminate catch 22s between different administrative bodies |
| (Peeters & Widlak, 2023) | **- ICT system standardization**<br>**- Opaque decision characteristics of ADM systems**<br>**- Historical and other biases in data**<br>**- Lack of transparency and accountability** | - Include mechanisms to control data quality and reliability<br>- Monitor the effects on primary processes and individual administrative decision-making<br>- Organize feedback mechanisms to the policy level<br>- Correct or override unintentional or disproportional consequences for individual cases |
| (Pel, 2022) | **- Data & Information silos obstruct causality and impact tracking**<br>**- Narrow focus on technical artefacts**<br>**- Linear rather than iterative**<br>**- Perverse behavioural incentives for administrative bodies** | - Improving system state visibility, through shortening feedback loops<br>- Improve baises for normative guidelines and there enforcement<br>- Mutually improve mental model of different organisational bodies<br>- Impose clearly defined system boundaries |
| (Ranerup & Svensson, 2023) | - Lack of legal and normative guidelines<br>- Lack of transparency and accountability<br>- Power of none-government system designers (privatization)<br>- Lack of education on the effects and risks of ADM system in users and managers | - Accomplishing personalization and client involvement as well as counteracting the digital divide |

| Article | Perpetuating Socio-technial Factors | Essential Design Advise |
|---|---|---|
| (van Engers & de Vries, 2019) | - Lack of transparency and accountability<br>- Opaque decision characteristics of ADM systems<br>- Historical and other biases in data<br>- Lack of participatory design methodology<br>- Lack of legal and normative guidelines | - Public participation in the development and implementation of AI algorithms |
| (van Noordt & Misuraca, 2020) | - Lack of transparency and accountability<br>- Lack of participatory design methodology<br>- Lack of risk and impact assessment framework | - More transparency and accountability through organisational design |
| (Veale & Brass, 2019) | - Lack of transparency and accountability<br>- Lack of participatory design methodology<br>- Historical and other biases in data<br>- Opaque decision characteristics of ADM systems<br>- Data & Information silos obstruct causality and impact tracking | - ADM systems should be accompanied by institutions, artefacts and strategies to monitor and evaluate their respective performance<br>- Design environment has to encourage professionalism |
| (Veale et al., 2018) | - Lack of transparency and accountability<br>- Historical and other biases in data<br>- Lack of participatory design methodology<br>- Lack of legal and normative guidelines | - ADM systems must be designed and studied within their environmental context<br>- Administrators, especially decision makers must be trained in ML&AI<br>- Iterative design is crucial to overcome accidents deployment |
| (Weyerer & Langer, 2019) | - Historical and other biases in data (pre-processing)<br>- In-processing biases, such as aggregation bias<br>- Post-processing biases through misuse of ADM tool<br>- Uneven power structures in the design of algorithmic | - Implement discrimination-related risk management and safety measures<br>-Development of strict and fair decision rules of AI applications that limit any conclusion based on non-relevant personal characteristics<br>- ADM systems should be avoided to hold people responsible and retain traceable judgments |
| (Widlak & Peeters, 2020) | - Lack of transparency and accountability<br>- Historical and other biases in data<br>- Data & Information silos obstruct causality and impact tracking<br>- Legal contamination via ICT<br>- ICT eliminating street level discretion<br>- Flawed incentive structure | - Burden of correcting bureaucratic ADM mistakes can not fall on the citizens |
| (Wirtz et al., 2021) | - Bureaucratic inertia<br>- Lack of transparency and accountability<br>- Technological lock-in<br>- Lack of legal and normative guidelines | - ADM systems will inadvertently create hazardous system states<br>- Strategies for mitigating resulting harms and accidents must be included into the design |
| (Wirtz & Müller, 2019) | - Lack of transparency and accountability<br>- Lack of legal and normative guidelines<br>- Lack of participatory design methodology<br>- Lack of risk and impact assessment framework | - Calls for holistic and integrated approach to AI governance, with a focus on stakeholder engagement and the need for clear policies and guidelines |
| (Wirtz et al., 2020) | - Lack of transparency and accountability<br>- Lack of legal and normative guidelines<br>- Lack of participatory design methodology<br>- Lack of risk and impact assessment framework | - Design needs to entail ongoing monitoring and evaluation of AI applications in public administration |

| Article | Perpetuating Socio-technial Factors | Essential Design Advise |
|---|---|---|
| (Young et al., 2019) | - Lack of transparency and accountability<br>- Histocial and other biases in data<br>- Lack of legal and normative guidelines<br>- Lack of participatory design methodology<br>- Lack of risk and impact assessment framework | - A holistic approach must consider the context of AI adoption, the design and implementation of AI systems, and then outcomes and impact of AI adoption |
| (Young et al., 2021) | - Lack of transparency and accountability<br>- Technocratic systems design approach<br>- Historical and other biases in data<br>- Uneven power structures in the design of algorithmic<br>- Lack of democratic participation and diversity in design<br>- Opacity of ADM systems makes challenging and understanding decisions difficult<br>- Power of none-government system designers (privatization) | - A holistic design approach considering technical design, institutional context, social impact, and ethical constraints is vital in order to alleviate harms |

# B

# Code & Result from Document RAG

In this appendix the code used to for the retrival-augmentation-generation document review procedure is shown. Hereinafter, the results from this process that was utilized in this research is shown. Several questions were posed yet not included, because they did not hold up to fact-checking. RAG is susceptible to model hallucination, therefore the source text was always retrieved as well, to verify the LLMs answers.

The model was questioned with a number of questions, here are several examples:

- How was FSV used inside of the Belastingdienst?
- Describe the process of deployment for the FSV, what were organisaations/departments responsibilities?
- What what were the responsibilities of the intensief toezichtteam?
- What were the responsibilities of the intensief toezichtteam?
- Please list important entities, individuals and Groups involved in the toeslagenaffair.
- What was the most impactul error that occured, which lead to the toeslagenaffair and citizen being harmed?
- What was the most impactul error that occured, which lead to the toeslagenaffair and citizen being harmed?
- What was the process of deployment followed for implementing FSV?
- Which workflow did the intensief toezichtteam follow and why was is flawed?
- Which workflow did the intensief toezichtteam follow and why was is flawed?
- Outline the decision process that would take place during toeslagen risk analysis, FSV, be precise.
- What happend when Opzet or Grove Schuld was established, how would the cases be handeld?
- Who was responsible for making decision to create a debt claim against citizens and who was responsible to follow up on this debt claim?
- What external organisation was contracted with developing the FSV for the Belastingdienst?
- What was the role of the landelike incasso centrum in the CAP?
- Who was the buisness owner of the FSV system?
- What company did the technical development of FSV?
- How many use cases did the FSV system cover?
- What was the role of the Datafundamenten en Analytics in managing risk models for the Toeslagen toezicht process?
- Who was the product owner of the FSV system?

Hereinafter the Bibliography for the documents the RAG model utilized within its answers:

2017/095 Belastingdienst treft 232 gezinnen met onevenredig harde actie | Nationale ombudsman. (2017, August 9). https://www.nationaleombudsman.nl/publicaties/onderzoeken/2017095-belastingdienst-treft-232-gezinnen-met-onevenredig-harde-actie

Anoniem. (n.d.). Parlementaire enquêtecommissie Fraudebeleid en Dienstverlening [Text]. Retrieved 29 January 2024, from https://www.tweedekamer.nl/kamerleden en commissies/commissies/pefd

Ashby, R. W. (n.d.). Instructie FSV. Archivering documenten Toezichtlijst en Beheer Uitsluitingen TVS [Text]. Retrieved 31 January 2024, from https://www.tweedekamer.nl/kamerstukken/detail

Boevenstreken van de Belastingdienst: De toeslagenaffaire in vogelvlucht. (2019, December 23). RTL Nieuws. https://www.rtlnieuws.nl/nieuws/artikel/4964351/toeslagenaffaire-belastingdienst-menno-snel-kinderopvang

Capgemini. (2013). Ontwikkelspecificatie Dagboek Fraude Signalering Voorziening FSV SOS NET Dienstverlening Belastingdienst versie 1.1. Contouren tegemoetkomingsregeling. (n.d.). Retrieved 30 January 2024, from https://www.staten-generaal.nl/9370000/1/j4nvjlhjvvt9eu4 j9vvkfvj6b325az/vlw0kkb2gmzf

Fenger, M., & Simonse, R. (n.d.). The implosion of the Dutch surveillance welfare state. Social Policy & Administration, n/a(n/a). https://doi.org/10.1111/spol.12998

Financiën, M. van. (2020, July 10). Aanbiedingsbrief FSV - Kamerstuk—Rijksoverheid.nl [Kamerstuk]. Ministerie van Algemene Zaken. https://doi.org/10/aanbiedingsbrief-fsv

Financiën, M. van. (2021a, September 29). Bijlage 1—Onderzoek PwC - Reconstructie en tijdlijn van het 'memo-Palmen'—Rapport—Rijksoverheid.nl [Rapport]. Ministerie van Algemene Zaken. https://www.rijksoverheid.nl/documenten/rap
1-onderzoek-pwc-reconstructie-en-tijdlijn-van-het-memo-palmen

Financiën, M. van. (2021b, September 29). Bijlage 2—PwC Reconstructie en tijdlijn van het 'memo-Palmen' Appendices (pagina's 1-301)—Rapport—Rijksoverheid.nl [Rapport]. Ministerie van Algemene Zaken. https://www.rijksoverheid.nl/documenten/rapporten/2021/09/29/bijlage-2—pwc-reconstructie-en-tijdlijn-van-het-memo-palmen-appendices-paginas-1-301

Financiën, M. van. (2021c, September 29). Bijlage 3—PwC Reconstructie en tijdlijn van het 'memo-Palmen' Appendices (pagina's 302-717)—Rapport—Rijksoverheid.nl [Rapport]. Ministerie van Algemene Zaken. https://www.rijksoverheid.nl/c
3-pwc-reconstructie-en-tijdlijn-van-het-memo-palmen-appendices-paginas-302-717

Financiën, M. van. (2021d, September 30). Aanbiedingsbrief onderzoek PwC reconstructie en tijdlijn van het memo van de vaktechnisch coördinator—Kamerstuk—Rijksoverheid.nl [Kamerstuk]. Ministerie van Algemene Zaken. https://www.rijksoverheid.nl/documenten/kamerstukken/2021/09/30/aanbiedingsbrief-onderzoek-pwc-reconstructie-en-tijdlijn-van-het-memo-van-de-vaktechnisch-coordinator

Financiën, M. van. (2021e, December 22). Onderzoek effecten FSV Particulieren—Rapport—Rijksoverheid.nl [Rapport]. Ministerie van Algemene Zaken. https://www.rijksoverheid.nl/documenten/rapporten/2021/12/22/bijlage-1-rapport-pwc-effecten-fsv-particulieren

Financiën, M. van. (2022a, January 19). Onderzoek Gegevensdeling met Derden—Rapport—Rijksoverheid.nl [Rapport]. Ministerie van Algemene Zaken. https://www.rijksoverheid.nl/documenten/rapporten/2022/01/19/bijlage-2-rapport-pwc-externe-gegevensdeling-uit-fsv

Financiën, M. van. (2022b, March 17). Onderzoek effecten FSV MKB - Rapport—Rijksoverheid.nl [Rapport]. Ministerie van Algemene Zaken. https://www.rijksoverheid.nl/documenten/rapporten/2022/03/17/onderzoek-effecten-fsv-mkb

Financiën, M. van. (2022c, May 30). Kamerbrief over Fraudesignaleringsvoorziening en vraagstuk institutioneel racisme—Kamerstuk—Rijksoverheid.nl [Kamerstuk]. Ministerie van Algemene Zaken. https://www.rijksoverheid.nl/documenten/ka
reactie-op-verzoeken-over-fraudesignaleringsvoorziening

Haag, D. (n.d.-a). BRIEF VAN DE PARLEMENTAIRE ONDERVRAGINGSCOMMISSIE. Haag, D. (n.d.-b). BRIEF VAN DE TIJDELIJKE COMMISSIE UITVOERINGSORGANISATIES. Kamerstuk 31066, nr. 992 | Overheid.nl > Officiële bekendmakingen. (n.d.). Retrieved 30 January 2024, from https://zoek.officielebekendmakingen.nl/kst-31066-992.html

KPMG. (2020). Rapportage verwerking van risicosignalen voor toezicht. 10.

Lijst van vragen en antwoorden over rapporten PwC over Fraudesignaleringsvoorziening (FSV)—Particulieren en externe gegevensdeling (Kamerstuk 31066-957) en over de reactie op verzoeken commissie over de rapporten van PwC over Fraudesignaleringsvoorziening (FSV)—Particulieren en externe gegevensdeling (Kamerstukken 31066-960)—Belastingdienst—EU monitor. (n.d.). Retrieved 14 February 2024, from https://www.eumonitor.eu/9353000/1/j9vvik7m1

Parlementair onderzoek uitvoeringsorganisaties (2020-2021). (n.d.). Retrieved 29 January 2024, from https://www.parlement.com onderzoek

PwC. (2021, November 30). Onderzoek effecten FSV Toeslagen—Rapport—Rijksoverheid.nl [Rapport]. Ministerie van Algemene Zaken. https://www.rijksoverheid.nl/documenten/rapporten/2021/12/03/onderzoek-pwc-effecten-fsv-toeslagen

Staten-Generaal, T. K. der. (2023, May 17). Belastingdienst; Brief regering; FSV gerelateerde onderwerpen [Officiële publicatie]. https://zoek.officielebekendmakingen.nl/kst-31066-1227.html

Tweete Kammer. (2022, February 20). Belastingdienst [Text]. https://www.tweedekamer.nl/kamerstukken/brieven regering/detail

Veiligheid, M. van J. en. (2022a, June 28). Brief Deelpublicatie CBS kenmerken gedupeerde gezinnen kinderopvangtoeslagenaffaire—Brief—Inspectie Justitie en Veiligheid [Brief]. Ministerie van Justitie en Veiligheid. https://www.inspectie-jenv.nl/Publicaties/brieven/2022/06/28/brief-deelpublicatie-cbs-kenmerken-gedupeerde-gezinnen-kinderopvangtoeslagenaffaire

Veiligheid, M. van J. en. (2022b, November 1). Brief Bevindingen Inspectie JenV onderzoek naar kinderopvangtoeslagaffaire en jeugdbescherming—Brief—Inspectie Justitie en Veiligheid [Brief]. Ministerie van Justitie en Veiligheid. https://www.inspectie-jenv.nl/Publicaties/brieven/2022/11/01/brief-bevindingen-inspectie-jenv-onderzoek-naar-kinderopvangtoeslagaffaire-en-jeugdbescherming

Veiligheid, M. van J. en. (2023, September 13). Inspectierapport Het kind van de rekening—Rapport—Inspectie Justitie en Veiligheid [Rapport]. Ministerie van Justitie en Veiligheid. https://www.inspectie-jenv.nl/Publicaties/rapporten/2023 het-kind-van-de-rekening

```
!pip install python-dotenv
```

```
from dotenv import load_dotenv
import os

load_dotenv()
```

```
YOUR_OPENAI_KEY= ''#input API key here, for privacy reasons the researchers key has been removed
YOUR_WEAVIATE_KEY= '' #input API key here, for privacy reasons the researchers key has been removed
YOUR_WEAVIATE_CLUSTER= 'https://doc-rag-i7qbhr8z.weaviate.network'#https://toeslag-3wt5sdfh.weaviate.network
```

## 0. Install Dependencies

```
!pip install langchain
!pip install weaviate-client
!pip install openai
!pip install unstructured
!pip install "unstructured[pdf]"
```

```
!apt-get install poppler-utils
```

```
!pip install tiktoken
```

## 1. Data Reading

```
from google.colab import drive
drive.mount('/content/drive')
```

```
directory_path = '/content/drive/My Drive/ColabNotebooks/Docs'
```

```
from langchain.document_loaders import DirectoryLoader

loader = DirectoryLoader(directory_path, glob="**/*.pdf")
data = loader.load()
```

```
print(f'You have {len(data)} documents in your data')
print(f'There are {len(data[0].page_content)} characters in your document')
```

## 2. Text Splitting

```
from langchain.text_splitter import RecursiveCharacterTextSplitter

text_splitter = RecursiveCharacterTextSplitter(chunk_size=1000, chunk_overlap=100)
docs = text_splitter.split_documents(data)
```

## 3. Embedding Conversion

```
from langchain.embeddings.openai import OpenAIEmbeddings

embeddings = OpenAIEmbeddings(openai_api_key = YOUR_OPENAI_KEY)
```

## 4. Vector Database Storage

```
import weaviate
from langchain.vectorstores import Weaviate

# connect Weaviate Cluster
auth_config = weaviate.AuthApiKey(api_key=YOUR_WEAVIATE_KEY)

WEAVIATE_URL = YOUR_WEAVIATE_CLUSTER
client = weaviate.Client(
    url=WEAVIATE_URL,
    additional_headers={"X-OpenAI-Api-Key": YOUR_OPENAI_KEY},
    auth_client_secret=auth_config,
    startup_period=10
)


# defined input structure
client.schema.delete_all()
client.schema.get()
schema = {
    "classes": [
        {
            "class": "Chatbot",
            "description": "Documents for chatbot",
            "vectorizer": "text2vec-openai",
            "moduleConfig": {"text2vec-openai": {"model": "ada", "type": "text"}},
            "properties": [
                {
                    "dataType": ["text"],
                    "description": "The content of the paragraph",
                    "moduleConfig": {
                        "text2vec-openai": {
                            "skip": False,
                            "vectorizePropertyName": False,
                        }
                    },
                    "name": "content",
                },
            ],
        },
    ]
}

client.schema.create(schema)

vectorstore = Weaviate(client, "Chatbot", "content", attributes=["source"])
```

```
# load text into the vectorstore
text_meta_pair = [(doc.page_content, doc.metadata) for doc in docs]
texts, meta = list(zip(*text_meta_pair))
vectorstore.add_texts(texts, meta)
```

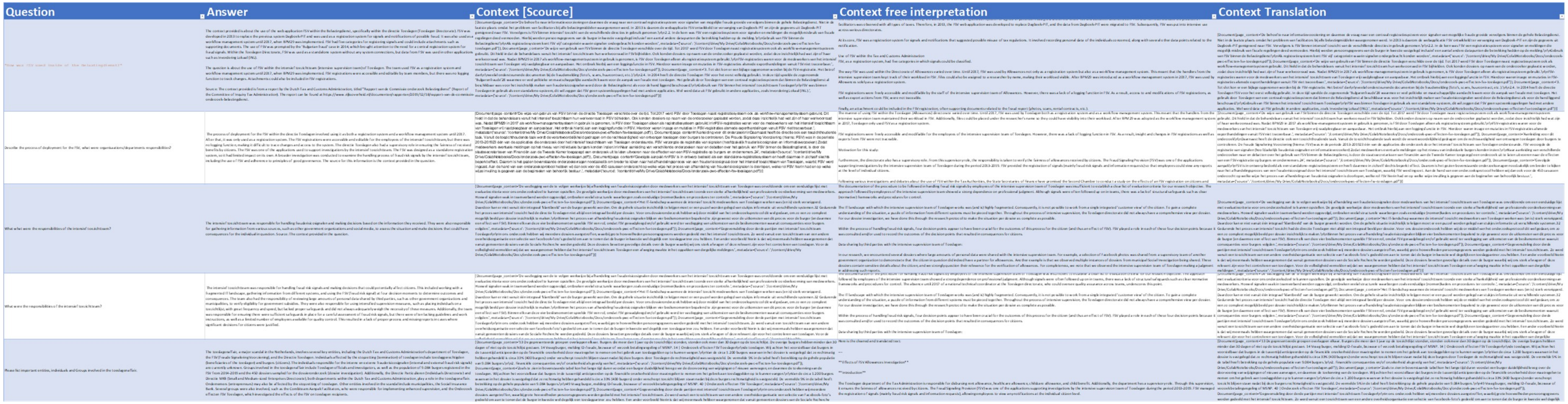
## 5. Similarity Search

```
query = "User Question: [...]"

# retrieve text related to the query
docs = vectorstore.similarity_search(query, k=20)
print(docs)
```

```
[Document(page_content='Op 28 april1 jl. hebben wij toegezegd u nader te informeren over de Fraude Signalering Voorziening (FSV) en het
```

## 6.Our Custom ChatBot

```
from langchain.chains.llm import LLMChain
from langchain.prompts import PromptTemplate
from langchain.chains.question_answering import load_qa_chain
from langchain.llms import OpenAI

from langchain.chains import MapReduceDocumentsChain, ReduceDocumentsChain
from langchain.text_splitter import CharacterTextSplitter
from langchain.chains.combine_documents.stuff import StuffDocumentsChain

llm = OpenAI(openai_api_key = YOUR_OPENAI_KEY,temperature=0)

# Map
map_template = """Follow exactly those 2 steps:
1. Read the context below and aggregrate this data
Context: {docs}
2. Answer the question using only this context, User Question: [...]?

Please provide your answer in English.
"""
map_prompt = PromptTemplate.from_template(map_template)
map_chain = LLMChain(llm=llm, prompt=map_prompt)


# Reduce
reduce_template = """The following is set of summaries:
{docs}
Take these and combine them into a final, insightful Answer of the question, User Question: What external organisation was contracted with d
Helpful Answer:"""
reduce_prompt = PromptTemplate.from_template(reduce_template)


# Run chain
reduce_chain = LLMChain(llm=llm, prompt=reduce_prompt)

# Takes a list of documents, combines them into a single string, and passes this to an LLMChain
combine_documents_chain = StuffDocumentsChain(
    llm_chain=reduce_chain, document_variable_name="docs", verbose=True
)

# Combines and iteratively reduces the mapped documents
reduce_documents_chain = ReduceDocumentsChain(
    # This is final chain that is called.
    combine_documents_chain=combine_documents_chain,
    # If documents exceed context for `StuffDocumentsChain`
    collapse_documents_chain=combine_documents_chain,
    # The maximum number of tokens to group documents into.
    token_max=2500,
)
# Combining documents by mapping a chain over them, then combining results
map_reduce_chain = MapReduceDocumentsChain(
    # Map chain
    llm_chain=map_chain,
    # Reduce chain
    reduce_documents_chain=reduce_documents_chain,
    # The variable name in the llm_chain to put the documents in
    document_variable_name="docs",
    # Return the results of the map steps in the output
    return_intermediate_steps=False,
)

text_splitter = CharacterTextSplitter.from_tiktoken_encoder(
    chunk_size=1000, chunk_overlap=0
)
split_docs = text_splitter.split_documents(docs)


map_reduce_chain.run(docs)


print(map_reduce_chain.run(split_docs))
```



| Question | Answer | Context [Scource] | Context free interpretation | Context Translation |
|---|---|---|---|---|
| "How was FSV used inside of the Belastingdienst?" | The context provided is about the use of the web application FSV within the Belastingdienst, specifically within the directie Toeslagen (Toeslagen Directorate). FSV was developed in 2013 to replace the previous system Dagboek-PIT and was used as a registration system for signals and notifications of possible fraud. It was also used as a workflow management system until 2017, when BPM29 was implemented. FSV had five categories for registering signals and could include attachments such as supporting documents. The use of FSV was prompted by the "Bulgarian fraud" case in 2014, which brought attention to the need for a central registration system for fraud signals. Within the Toeslagen Directorate, FSV was used as a standalone system without any system connections, but data from FSV was used in other applications such as Invordering Lokaal (INL).<br>The question is about the use of FSV within the intensief toezichtteam (intensive supervision team) of Toeslagen. The team used FSV as a registration system and workflow management system until 2017, when BPM29 was implemented. FSV registrations were accessible and editable by team members, but there was no logging function to track changes. Attachments could also be included in FSV registrations.<br>Source: The context provided is from a report by the Dutch Tax and Customs Administration, titled "Rapport van de Commissie onderzoek Belastingdienst" (Report of the Committee of Inquiry Tax Administration). The report can be found at https://www.rijksoverheid.nl/documenten/rapporten/2019/12/18/rapport-van-de-commissie-onderzoek-belastingdienst. | [illegible] | facilitators was observed with all types of taxes. Therefore, in 2013, the FSV web application was developed to replace Dagboek-PIT, and the data from Dagboek-PIT were migrated to FSV. Subsequently, FSV was put into intensive use across various directorates.<br>At its core, FSV was a registration system for signals and notifications that suggested possible misuse of tax regulations. It involved recording personal data of the individuals concerned, along with several other data points related to the notification.<br>Use of FSV within the Tax and Customs Administration<br>FSV, as a registration system, had five categories in which signals could be classified.<br>The way FSV was used within the Directorate of Allowances varied over time. Until 2017, FSV was used by Allowances not only as a registration system but also as a workflow management system. This meant that the handlers from the intensive supervision team kept track of their workload in FSV. Files could also be assigned to a researcher by name, making their workload visible. After BPM29 was introduced as a workflow management system in 2017, FSV was used by Allowances solely as a registration system.<br>FSV registrations were freely accessible and modifiable by the staff of the intensive supervision team of Allowances. However, there was a lack of a logging function in FSV. As a result, access to and modifications of FSV registrations, as well as export actions from FSV, were not traceable.<br>Finally, an attachment could be included in the FSV registration, often supporting documents related to the fraud report (photos, scans, rental contracts, etc.). | [illegible] |
| Describe the process of deployment for the FSV, what were organisations/departments responsibilities? | The process of deployment for the FSV within the directie Toeslagen involved using it as both a registration system and a workflow management system until 2017. After that, it was only used as a registration system. The FSV registrations were accessible and editable for the employees of the intensief toezichtteam, but there was no logging function, making it difficult to trace changes and access to the system. The directie Toeslagen also had a supervisory role in ensuring the fairness of received benefits by citizens. The FSV was one of the applications used to support investigations by the intensief toezichtteam. The FSV was designed as a standalone registration system, so it had limited impact on its own. A broader investigation was conducted to examine the handling process of fraud risk signals by the intensief toezichtteam, including the use of FSV and adherence to principles of good governance. The source for this information is the context provided in the question. | [illegible] | The manner of using FSV within the Toeslagen (Allowances) directorate varied over time. Until 2017, FSV was used by Toeslagen both as a registration system and as a workflow management system. This meant that the handlers from the intensive supervision team maintained their workload in FSV. Additionally, files could be placed under the researcher's name so they could have visibility into their workload. After BPM 29 was adopted as the workflow management system in 2017, Toeslagen only used FSV as a registration system.<br>FSV registrations were freely accessible and modifiable for the employees of the intensive supervision team of Toeslagen. However, there was a lack of logging function in FSV. As a result, insight and changes in FSV registrations as well as exports from FSV were not traceable.<br>Motivation for this study:<br>Furthermore, the directorate also has a supervisory role. From this supervisory role, the responsibility is taken to verify the fairness of allowances received by citizens. The Fraud Signaling Provision (FSV) was one of the applications supporting investigations by the intensive supervision team of Toeslagen during the period 2013-2019. FSV provided the registration of signals (mainly fraud risk signals and information requests) so that employees could view any reports at the level of individual citizens.<br>Following various investigations and debates about the use of FSV within the Tax Authorities, the State Secretaries of Finance have promised the Second Chamber to conduct a study on the effects of an FSV registration on citizens and | [illegible] |
| What what were the responsibilities of the intensief toezichtteam? | The intensief toezichtteam was responsible for handling frauderisicosignalen and making decisions based on the information they received. They were also responsible for gathering information from various sources, such as other government organizations and social media, to assess the situation and make decisions that could have consequences for the individual in question. Source: The context provided in the question. | [illegible] | The documentation of the procedure to be followed in handling fraud risk signals by employees of the intensive supervision team of Toeslagen was insufficient to establish a clear list of evaluation criteria for our research objective. The approach followed by employees of the intensive supervision team showed a strong dependence on professional judgment. Although signals were often followed up on in teams, there was a lack of structural safeguards such as clear (normative) frameworks and procedures for control.<br>The IT landscape with which the intensive supervision team of Toeslagen works was (and is) highly fragmented. Consequently, it is not possible to work from a single integrated 'customer view' of the citizen. To gain a complete understanding of the situation, a puzzle of information from different systems must be pieced together. Throughout the process of intensive supervision, the Toeslagen directorate did not always have a comprehensive view per dossier. For our dossier investigation, we have done this through the research protocol to make the situation per dossier as complete as possible.<br>Within the process of handling fraud risk signals, four decision points appear to have been crucial for the outcomes of this process for the citizen (and thus an effect of FSV). FSV played a role in each of these four decision points because it was consulted and/or used to record the outcomes of the decision points that resulted in consequences for citizens.<br>Data sharing by third parties with the intensive supervision team of Toeslagen:<br>In our research, we encountered several dossiers where large amounts of personal data were shared with the intensive supervision team. For example, a selection of Facebook photos was shared from a supervisory team of another government organization to demonstrate that the citizen in question did indeed have a partner for allowances. Another example is that we observed multiple instances of dossiers from municipal Social Investigation being shared. These dossiers contain sensitive details about the citizen, and we strongly question their relevance for the verification of allowances. For completeness, we note that we observed the intensive supervision team of Toeslagen making a judgment in addressing such reports. | [illegible] |
| What were the responsibilities of the intensief toezichtteam? | The intensief toezichtteam was responsible for handling fraud risk signals and making decisions that could potentially affect citizens. This included working with a fragmented IT landscape, gathering information from different systems, and using the FSV (fraud risk signal) at four decision moments to determine outcomes and consequences. The team also had the responsibility of reviewing large amounts of personal data shared by third parties, such as other government organizations and municipalities, to verify eligibility for government subsidies. They were also responsible for using intensified supervision measures, such as placing individuals on a toezichtlijst, with great frequency and speed, but lacked proper safeguards and did not always adequately weigh the necessity of these measures. Additionally, the team was responsible for ensuring there were sufficient safeguards in place for a careful assessment of fraud risk signals, but there were often lacking guidelines and work instructions, as well as a limited number of employees available for quality control. This resulted in a lack of proper process and missing reports in cases where significant decisions for citizens were justified. | [illegible] | The documentation of the procedure for handling fraud risk signals by employees of the intensive supervision team of Toeslagen was insufficient to establish a clear list of evaluation criteria for our research objective. The approach followed by employees of the intensive supervision team showed a strong dependence on professional judgment. Although signals were often followed up on in teams, there was a lack of structural safeguards such as clear normative frameworks and procedures for control. The absence until 2017 of a national technical coordinator at the Toeslagen directorate, who could oversee quality assurance across teams, underscores this point.<br>The IT landscape with which the intensive supervision team of Toeslagen works was (and is) highly fragmented. Consequently, it is not possible to work from a single integrated 'customer view' of the citizen. To gain a complete understanding of the situation, a puzzle of information from different systems must be pieced together. Throughout the process of intensive supervision, the Toeslagen directorate did not always have a comprehensive view per dossier. For our dossier investigation, we have done this through the research protocol to make the situation per dossier as complete as possible.<br>Within the process of handling fraud risk signals, four decision points appear to have been crucial for the outcomes of this process for the citizen (and thus an effect of FSV). FSV played a role in each of these four decision points because it was consulted and/or used to record the outcomes of the decision points that resulted in consequences for citizens.<br>Data sharing by third parties with the intensive supervision team of Toeslagen: | [illegible] |
| Please list important entities, individuals and Groups involved in the toeslagenaffair. | The toeslagenaffair, a major scandal in the Netherlands, involves several key entities, including the Dutch Tax and Customs Administration's department of Toeslagen, the FSV (Fraude Signalering Voorziening), and the Directie Toeslagen. Individuals affected by the stopzetting (termination) of toeslagen include toeslaggerechtigden (beneficiaries of the toeslagen) and burgers (citizens). The individuals responsible for the interne en externe frauderisicosignalen (internal and external fraud risk signals) are currently unknown. Groups involved in the toeslagenaffair include Toeslagen officials and investigators, as well as the population of 9.084 burgers registered in the FSV from 2014-2019 and the 450 dossiers sampled for the dossieronderzoek (dossier investigation). Additionally, the Directie Particulieren (Individuals Directorate) and Directie MKB (Small and Medium-sized Enterprises Directorate), both departments within the Dutch Tax and Customs Administration, play a role in the toeslagenaffair. Ondernemers (entrepreneurs) may also be affected by the stopzetting of toeslagen. Other entities involved in the scandal include municipalities, the Social Insurance Bank. Several groups were also involved, such as the Combiteam Aanpak Facilitation, who were responsible for implementing enhanced supervision, and the Onderzoek effecten FSV Toeslagen, which investigated the effects of the FSV on toeslagen recipients. | [illegible] | Here is the cleaned and translated text:<br>---<br>**Effects of FSV Allowances Investigation**<br>**Introduction**<br>The Toeslagen department of the Tax Administration is responsible for disbursing rent allowance, healthcare allowance, childcare allowance, and child benefit. Additionally, the department has a supervisory role. Through this supervision, it ensures the fairness of allowances received by citizens. The Fraud Signaling Provision (FSV) was one of the applications supporting investigations by the intensive supervision team of Toeslagen during the period 2013-2019. FSV managed the registration of signals (mainly fraud risk signals and information requests), allowing employees to view any notifications at the individual citizen level. | [illegible] |

Figure B.1: Example of question documentation of GPT based RAG-model

# C

# Safety Control Diagram & Structure

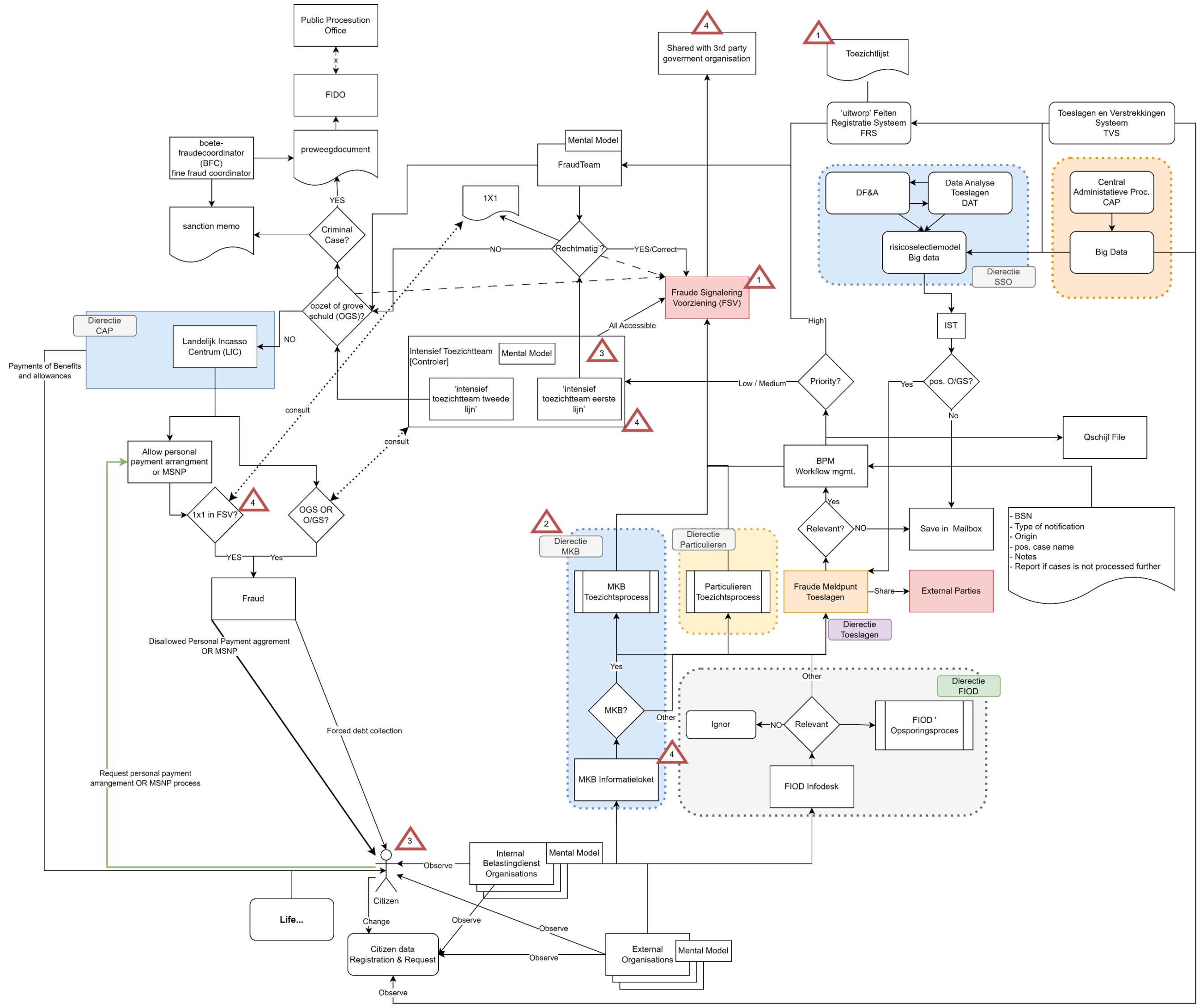

Public Procesution Office
x
FIDO
preweegdocument
boete-fraudecoordinator (BFC) fine fraud coordinator
sanction memo
YES
Criminal Case?
opzet of grove schuld (OGS)?
NO
Diercetie CAP
Landelijk Incasso Centrum (LIC)
Payments of Benefits and allowances
consult
Allow personal payment arrangment or MSNP
1x1 in FSV?
4
OGS OR O/GS?
YES
Yes
Fraud
Disallowed Personal Payment aggrement OR MSNP
Forced debt collection
Request personal payment arrangement OR MSNP process
3
Citizen
Life...
Change
Citizen data Registration & Request
Observe
Internal Belastingdienst Organisations
Mental Model
Observe
Observe
Observe
External Organisations
Mental Model
Observe
1X1
Mental Model
FraudTeam
Rechtmatig'?
NO
YES/Correct
4
Shared with 3rd party goverment organisation
Fraude Signalering Voorziening (FSV)
1
All Accessible
Intensief Toezichtteam [Controler]
Mental Model
3
'intensief toezichtteam tweede lijn'
'intensief toezichtteam eerste lijn'
4
consult
2
Dierectie MKB
MKB Toezichtsprocess
Yes
MKB?
Other
4
MKB Informatieloket
Dierectie Particulieren
Particulieren Toezichtsprocess
1
Toezichtlijst
'uitworp' Feiten Registratie Systeem FRS
Toeslagen en Verstrekkingen Systeem TVS
DF&A
Data Analyse Toeslagen DAT
risicoselectiemodel Big data
Dierectie SSO
Central Administatieve Proc. CAP
Big Data
IST
High
Priority?
Low / Medium
pos. O/GS?
Yes
No
Qschijf File
BPM Workflow mgmt.
Yes
Relevant?
NO
Save in Mailbox
- BSN
- Type of notification
- Origin
- pos. case name
- Notes
- Report if cases is not processed further
Fraude Meldpunt Toeslagen
Share
External Parties
Dierectie Toeslagen
Other
Dierectie FIOD
Ignor
NO
Relevant
FIOD ' Opsporingsproces
FIOD Infodesk

Figure C.1: Process flow Toeslagenaffair

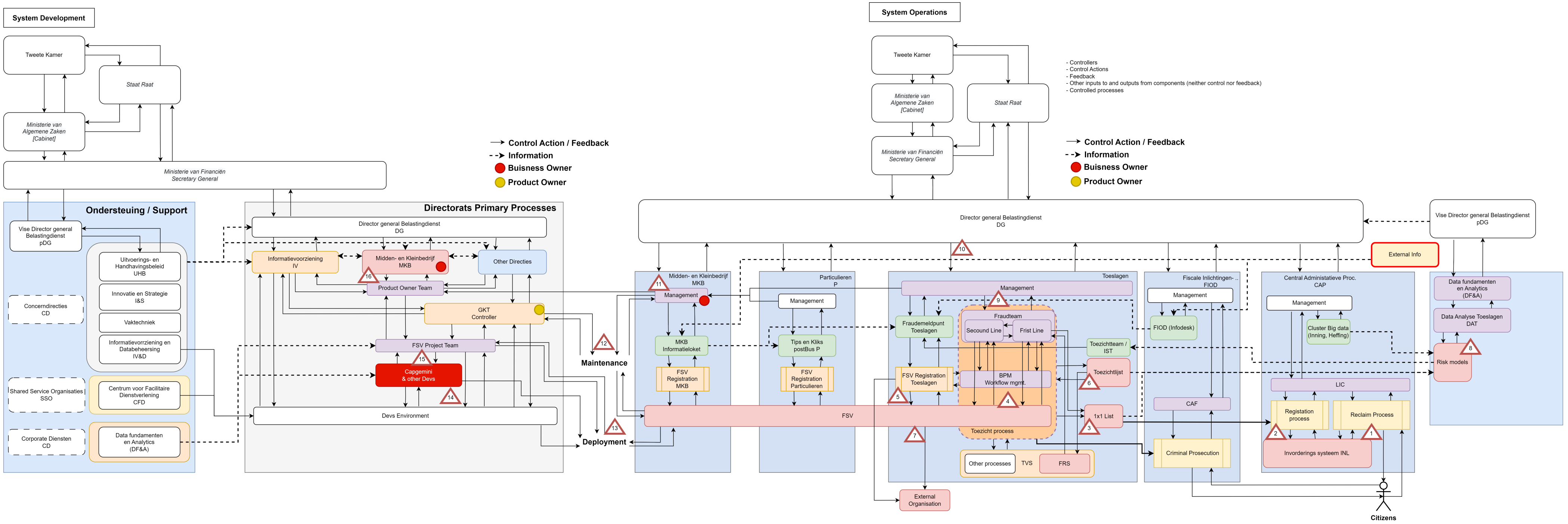

System Development
Tweete Kamer
Staat Raat
Ministerie van Algemene Zaken [Cabinet]
Ministerie van Financiën Secretary General
Ondersteuing / Support
Vise Director general Belastingdienst pDG
Uitvoerings- en Handhavingsbeleid UHB
Innovatie en Strategie I&S
Vaktechniek
Informatievorrziening en Databeheersing IV&D
Concerndirecties CD
Shared Service Organisaties SSO
Centrum voor Facilitaire Dienstverlening CFD
Corporate Diensten CD
Data fundamenten en Analytics (DF&A)
Control Action / Feedback
Information
Buisness Owner
Product Owner
Directorats Primary Processes
Director general Belastingdienst DG
Informatievoorziening IV
Midden- en Kleinbedrijf MKB
Other Directies
Product Owner Team
GKT Controller
FSV Project Team
Capgemini & other Devs
Devs Environment
Maintenance
Deployment
System Operations
Tweete Kamer
Ministerie van Algemene Zaken [Cabinet]
Staat Raat
Ministerie van Financiën Secretary General
- Controllers
- Control Actions
- Feedback
- Other inputs to and outputs from components (neither control nor feedback)
- Controlled processes
Control Action / Feedback
Information
Buisness Owner
Product Owner
Director general Belastingdienst DG
Vise Director general Belastingdienst pDG
External Info
Midden- en Kleinbedrijf MKB
Management
MKB Informatieloket
FSV Registration MKB
Particulieren P
Management
Tips en Kliks postBus P
FSV Registration Particulieren
Toeslagen
Management
Fraudemeldpunt Toeslagen
Fraudteam
Secound Line
Frist Line
BPM Workflow mgmt.
Toezicht process
FSV Registration Toeslagen
Toezichtteam / IST
Toezichtlijst
1x1 List
FSV
Other processes
TVS
FRS
External Organisation
Fiscale Inlichtingen- .. FIOD
Management
FIOD (Infodesk)
CAF
Criminal Prosecution
Central Administatieve Proc. CAP
Management
Cluster Big data (Inning, Heffing)
LIC
Registation process
Reclaim Process
Invorderings systeem INL
Data fundamenten en Analytics (DF&A)
Data Analyse Toeslagen DAT
Risk models
Citizens
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16

Figure C.2: Control structure Toeslagenaffair

| ID | Source | Type UCA | Type | Control Action / Feedback / Information | Context | System State |
|---|---|---|---|---|---|---|
| 1.1 | LIC does | Not provid | Control Action | Appeal process option | in cases were ”fraud/neglegance” or ”1x1” had been determined | Operations |
| 1.2 | LIC caseworkers | provid | Control Action | 100% debt reclaim & ”personal payment plan” rejection | in cases were ”fraud/neglegance” or ”1x1” had been determined | Operations |
| 1.3 | LIC Reclaim Process does | Not provid | Feedback | A report to the Toeslagen on case handeling | if they received information from toeslagen and the fraudteam | Operations |
| 2.1 | LIC registration | Misinterprets | Data/Other | ”1x1” list it receives by Toeslagen | during registration of incomming signals in LIC from toeslagen | Operations |
| 2.2 | INL system does | Not provid | Data/Other | The correct data structure to capture the complexity of the decision made by toeslagen | in cases were ”fraud/neglegance” or ”1x1” had been determined | Development |
| 3.1 | Toeslagen does | Not provid | Data/Other | Provide sufficant information regarding the ”1x1” decision | During registration of new cases in ”1x1” list and update of existing cases | Operations |
| 3.2 | Toeslagen does | Not provid | Control Action | A logging controler to understand and changes to instances in list over time | During registation or changes to existing cases in ”1x1” list | Operations |
| 3.3 | Toeslagen does | Not provid | Data/Other | A clear, documented and updated ontology of the data that is stored in ”1x1” | during systems deployment and operations over time | Operations |
| 3.4 | The Fraudteam | applied too long | Control Action | kept individuals on the ”1x1” | in cases were evidence against the contrary had surfaced or individuals had not exhibited ”suspicous behaviour” any longer | Operations |
| 3.5 | The Fraudteam did | Stopped too soon | Control Action | A scheduled revison and cleaning of ”1x1” lists | during systems operations over time | Operations |
| 4.1 | FSV | Not provid | Control Action | A logging function to trag decision and information changes over time | during systems operations over time | Operations |
| 4.2 | FSV | Not provid | Control Action | Function to ”clean” system in regular time intervals | during systems operations over time | Operations |
| 4.3 | FSV | Not provid | Data/Other | Sufficant data structure to handel the deversity of signals and there different implications | during systems operations over time | Development |
| 5.1 | Fraudemeldpunt Toeslagen did | Not provid | Control Action | Sufficant audit over FSV data registration process | during systems operations over time | Operations |
| 5.2 | FSV registration Toeslagen did | Not provid | Control Action | Oversight of other FSV signals relevant to current signal | During new system registration | Operations |
| 5.3 | Fraudemeldpunt Toeslagen | Not provid | Control Action | Sufficant risk signal verification | Prior to sharing signals with external parties | Operations |
| 5.4 | Fraudemeldpunt Toeslagen | Not provid | Data/Other | Enough information regarding signal interpretation and applicability | While sharing information with external parties | Operations |
| 6.1 | Fraudteam does | Not provid | Control Action | A logging controler to understand and changes to instances in list over time | During registation or changes to existing cases in ”1x1” list | Operations |
| 6.2 | Fraudteam does | Not provid | Data/Other | A clear, documented and updated ontology of the data that is stored in ”Toezichtlijst” | during systems deployment and operations over time | Operations |
| 6.3 | The Fraudteam | applied too long | Control Action | kept individuals on the ”Toezichtlijst” | in cases were evidence against the contrary had surfaced or individuals had not exhibited ”suspicous behaviour” any longer | Operations |
| 6.4 | The Fraudteam did | Stopped too soon | Control Action | A scheduled revison and cleaning of ”Toezichtlijst” lists | during systems operations over time | Operations |
| 7.1 | Fraudteam did | Not provid | Control Action | Sufficant risk signal verification | Prior to sharing signals with external parties out of FSV | Operations |
| 7.2 | FSV did | Not provid | Data/Other | Enough information regarding signal interpretation and applicability | While sharing information with external parties | Operations |
| 8.1 | DAT did | Not provid | Control Action | Transparent Version control of risk models in use | While utilizing different instances of model in operations | Operations |
| 8.2 | DAT did | Not provid | Data/Other | Inforamtion which model had identified specific signals | While utilizing different instances of model in operations | Operations |
| 8.3 | DAT did | Not provid | Control Action | Down stream audits of how models are utilized | During decision making by Fraudteam operators | Operations |
| 9.1 | Toeslagen Mgmt. | Not provid | Control Action | Safety audit requirements to Toezicht process | On a periodical bases | Operations |
| 9.2 | Toezicht team | Not provid | Feedback | To managemnt in a predetermined process | If safety/ADM model incursion were detected | Operations |
| 9.3 | Toeslagen Mgmt. | Provided | Control Action | Directive to increase output | After work backlock and the related workload had reported to be high | Operations |
| 9.2 | Toezicht team | Not provid | Feedback | To development team | On data quality and registration issues, use case deterioration | Development |
| 10.1 | Director General | Not provid | Control Action | Safety focused process Audits for Toeslagen process responsibilities | On a periodical baises | Operations |
| 10.2 | Toeslagen Mgmt. | Not provid | Feedback | To managemnt in a predetermined process | If safety/ADM model incursion were detected | Operations |
| 11.1 | MKB Mgmt. | Not provid | Feedback | To development team on FSV performance for all devisions | On a periodical baises | Development |
| 11.2 | MKB Mgmt. | Not provid | Control Action | Audits of other devisions to check if FSV is utilized as intended | On a periodical baises | Operations |
| 12.1 | System Maintenance | Not provid | Control Action | Continous monitoring of buisness functionality in FSV system, only focusing on technical factors | On a periodical baises | Development |
| 12.2 | System Maintenance | Not provid | Information | On data quality present in system | during continous operation | Operations |
| 12.3 | System Maintenance | Not provid | Feedback | To development regarding indentified business logic FSV system discrapencies | Continously or after issues had been raised | Development |
| 13.1 | System Development | Not provid | Control Action | Improve Buisness logic utility of FSV system | Continously or after issues had been raised | Development |
| 13.2 | System Development | Not provid | Control Action | Deploy system in controled, small batches | during system deployment | Development |
| 13.3 | System Development | Not provid | Control Action | Test a the holisitic process | during system deployment | Development |
| 13.4 | System Development | Not provid | Control Action | Data quality standards, through design | that would have helped aid system operators, in design phase | Development |
| 14.1 | System Development | Not provid | Control Action | An accurate design enviornemnt resembeling deployment environment | To system developers and third party developers during design and deployment phase | Development |
| 15.1 | FSV Project Team | Not provid | Control Action | Restricrtion on Use case and deployment size | during design | Development |
| 15.2 | FSV Project Team | Not provid | Control Action | holistic system requirements/tests, that would have required the buisness logic to be verified | during design | Development |
| 16.1 | MKB | Not provid | Control Action | A transparent process to incooperate or exclute other devision requirements | into the design process | Development |

| | | | | | | |
|---|---|---|---|---|---|---|
| 16.2 | MKB | Not provid | Data/Other | Clear tradoff implciation of design objectives determined by the buisness owner | during conceptual phase | Development |
| 16.3 | MKB | Not provid | Control Action | Incooperate a holistic system safety approach that required the development to make design assumptions explicit | during conceptual phase | Development |

# D

# Capgemini Use Case specification FSV 2013

# 2 Specificatie Functionaliteit

## 2.1 Use cases

Deze paragraaf geeft een opsomming van de smart use cases binnen de applicatie Dagboek FSV. Hierin staat alle functionaliteit die de applicatie zal krijgen.

| Use Case name | Complexity | Priority | Package |
|---|---|---|---|
| UC102 Zoeken alle signalen | 5 | 1 - Must have | Behandelen |
| UC100 AanmakenBehandelen Aangiftefraude | 4 | 1 - Must have | Behandelen |
| UC200 Aanmaken/Behandelen InformatieVerzoek | 4 | 1 - Must have | Behandelen |
| UC400 AanmakenBehandelen TipKlikMelding | 4 | 1 - Must have | Behandelen |
| UC500 AanmakenBehandelen Diversen_Project (registratie van..) | 4 | 1 - Must have | Behandelen |
| UC600 Raadplegen PIT | 4 | 1 - Must have | Klantendienst |
| UC601 Zoeken en Bepalen voorkomen in PIT | 3 | 1 - Must have | Klantendienst |
| UC100 AanmakenBehandelen Aangiftefraude | 2 | 1 - Must have | Behandelen |
| UC900 Importeren Aangiftefraudes | 8 | 1 - Must have | Behandelen |
| UC1012 Filter en toon alle signalen | 4 | 1 - Must have | Rapporteren |
| UC1002 Filter en toon Aangiftefraudes | 4 | 1 - Must have | Rapporteren |
| UC1011 Exporteren alle gegevens van alle signalen naar Excel | 4 | 1 - Must have | Rapporteren |
| UC1001 Exporteren alle gegevens Aangiftefraudes naar Excel | 5 | 1 - Must have | Rapporteren |
| UC905 Importeren DiversenProjecten | 8 | 1 - Must have | Behandelen |
| UC906 Massaal Bijwerken Aangiftefraudes | 8 | 1 - Must have | Behandelen |
| UC109 Toon aantal signalen Hoog BCA | 4 | 1 - Must have | Behandelen |
| UC110 Toon aantal signalen Rappel | 4 | 1 - Must have | Behandelen |
| UC411 AanmakenBewerken Aantekening TipKlikMelding | 8 | 1 - Must have | Behandelen |
| UC511 AanmakenBewerken Aantekening Project Overig | 8 | 1 - Must have | Behandelen |
| UC211 AanmakenBewerken Aantekening InformatieVerzoek | 8 | 1 - Must have | Behandelen |
| UC111 AanmakenBewerken Aantekening Aangiftefraude | 8 | 1 - Must have | Behandelen |
| UC1013 Mijn openstaande posten | 5 | 1 - Must have | Rapporteren |
| UC804 Beheren overige lijsten | 5 | 1 - Must have | Beheren |
| UC802 Beheren Competente Eenheid | 4 | 1 - Must have | Beheren |
| UC801 Zoek medewerker | 3 | 1 - Must have | Beheren |
| UC803 Zoeken Competente Eenheid | 4 | 1 - Must have | Beheren |
| UC800 Beheren medewerker | 4 | 1 - Must have | Beheren |
| UC105 Bijlage toevoegen | 10 | 2 - Should have | Behandelen |
| UC806 Zoeken regionaal project | 3 | 1 - Must have | Beheren |
| UC805 Beheren regionaal project | 4 | 1 - Must have | Beheren |
| UC112 Verwijderen bijlage | 10 | 1 - Must have | Beheren |
| UC808 Beheren Regio | 4 | 2 - Should have | Beheren |
| UC807 Selecteren overige lijsten | 3 | 1 - Must have | Beheren |
| UC606 Raadplegen bijlagen | 3 | 1 - Must have | Behandelen |
| UC607 Raadplegen Aantekening | 3 | 1 - Must have | Behandelen |
| UC403 Verwijderen TipKlikMelding | 3 | 2 - Should have | Behandelen |
| UC503 Verwijderen Diversen_Project | 3 | 2 - Should have | Behandelen |
| UC104 Verwijderen Aangiftefraude | 3 | 2 - Should have | Behandelen |
| UC404 Overzetten TipKlikMelding naar aangiftefraude | 5 | 2 - Should have | Behandelen |
| UC407 Overzetten TipKlikMelding naar project/overig | 5 | 2 - Should have | Behandelen |

| | | | |
|---|---|---|---|
| UC107 Dupliceren Aangiftefraude | 4 | 2 - Should have | Behandelen |
| UC505 Dupliceren Diversen_Project | 4 | 2 - Should have | Behandelen |
| UC1018 Exporteren Mijn openstaande posten | 5 | 2 - Should have | Rapporteren |
| UC1014 Exporteren Aantallen Aangiftefraude naar Excel | 4 | 1 - Must have | Rapporteren |
| UC1015 Filter en toon Aantallen Aangiftefraude | 4 | 1 - Must have | Rapporteren |
| UC1016 Exporteren aantallen BCA meldingen naar Excel | 5 | 1 - Must have | Rapporteren |
| UC1017 Filter en toon aantallen BCA meldingen | 5 | 1 - Must have | Rapporteren |
| UC106 Exporteer aangifte fraude naar Word | 8 | 2 - Should have | Behandelen |
| CR-1 | | | |
| CR-2 | | | |
| CR-3 | | | |

De complexiteit in bovenstaande tabel geeft een inschatting van de hoeveelheid werk dat verricht moet worden om de Smart use case te realiseren, uitgedrukt in een aantal punten. De totale omvang van de applicatie Dagboek FSV komt op een totaal van 225 smart use case punten, hierin zijn change requests niet meegerekend. De MoSCow prioriteiten van de use cases zijn ook aangegeven in de tabel. Voor de prioriteit geldt 1=Must Have, 2=Should Have, 3=Could Have en 4=Won't Have.
De applicatie is onderverdeeld in een aantal 'onderdelen' of 'packages'. Packages in de applicatie Dagboek FSV zijn behandelen, rapporteren en beheren. Behandelen omschrijft bijvoorbeeld functionaliteit waarbij de gebruiker signalen kan zoeken, raadplegen, muteren zoals bijvoorbeeld het opvoeren van een nieuw signaal. Per use case is aangegeven in welk applicatie onderdeel hij valt.

## 2.2 Applicatieonderdelen

De functionaliteit zal worden gerealiseerd in een webapplicatie.

## 2.3 Use Case Diagrammen

In deze paragraaf zijn de use case diagrammen opgenomen waaruit de samenhang van de hierboven opgesomde use cases blijkt. In de diagrammen geven de cijfers onderin de use cases de complexiteit van de use case weer.
Indien gesproken wordt over 'Onderhouden' wordt hieronder verstaan het inzien, toevoegen, wijzigen en verwijderen van een entiteit.

Hieronder volgt een korte beschrijving van de verschillende soorten relaties tussen use cases. Een relatie heeft altijd een richting, waardoor er een 'van' use case is en een 'naar' use case.

### Include relatie

Een 'include' relatie tussen twee use cases betekent dat de 'van' use case de 'naar' use case tijdens het normale verloop altijd zal aanroepen. De 'naar' use case bevat dus functionaliteit die nodig is om de 'van' use case uit te voeren. Voorbeeld: Om een signaal te kunnen verwijderen (UC104 Verwijderen Aangiftefraude) zal deze eerst opgezocht moeten worden (UC101 Zoeken aangifte fraude) en moet deze Include kan dus vertaald worden door 'roept altijd aan'. De pijl staat van de aanroepende use case naar de aangeroepen use case.

### Extend relatie

Een 'extend' relatie tussen twee use cases betekent dat de 'van' use case extra functionaliteit toevoegt aan de 'naar' use case. Bijvoorbeeld bij het behandelen van een signaal (UC100 AanmakenBehandelen Aangiftefraude) kan de gebruiker de gegevens naar word exportren (UC106 Exporteer aangifte fraude naar Word). De relatie ligt in dit geval van de (eventueel) aangeroepen use case naar de aanroepende use case, dus omgekeerd aan de 'include' relatie.

# E

# Improved Safety Control Structure

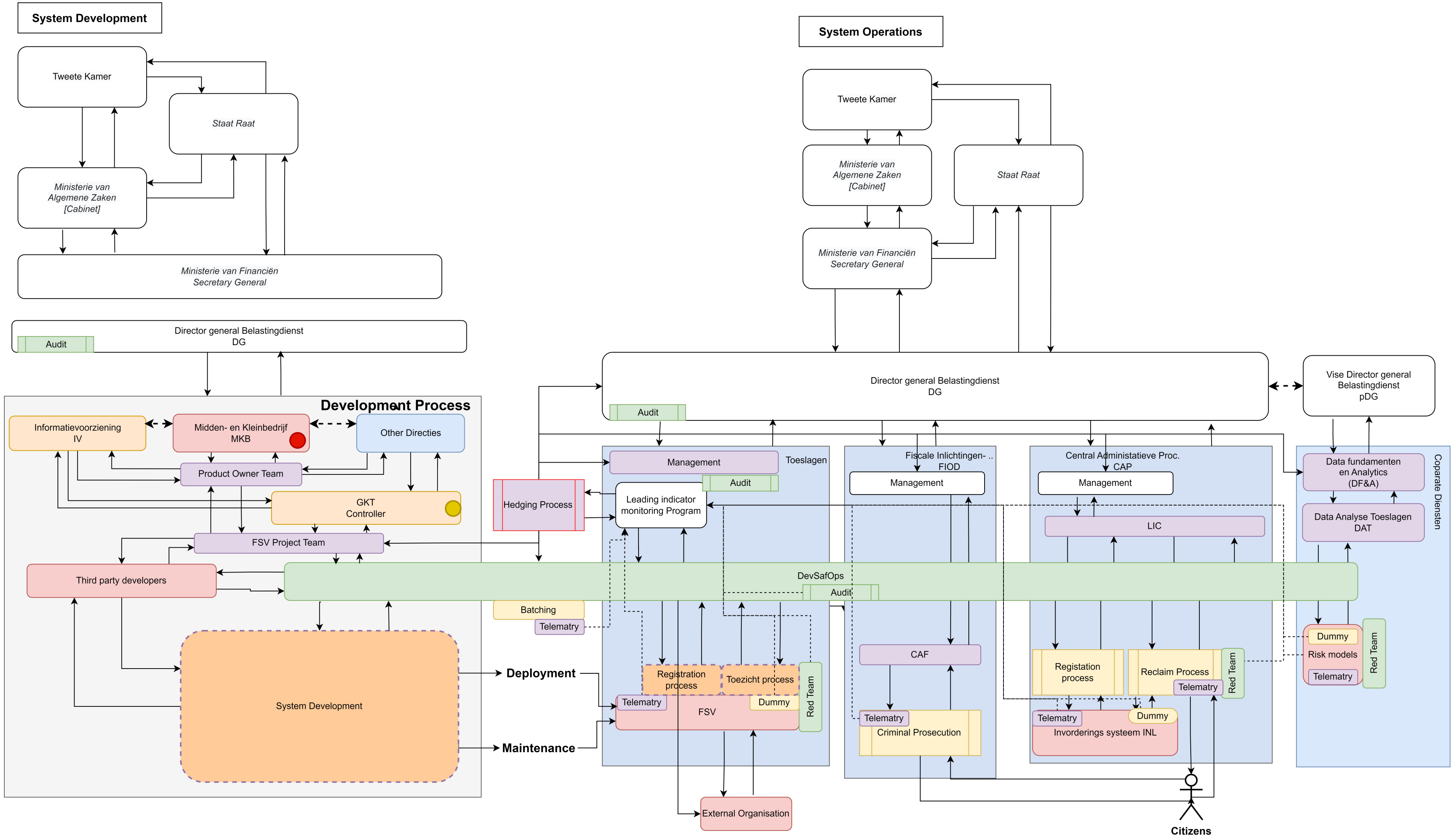


Figure E.1: Control structure Toeslagenaffair including Interventions