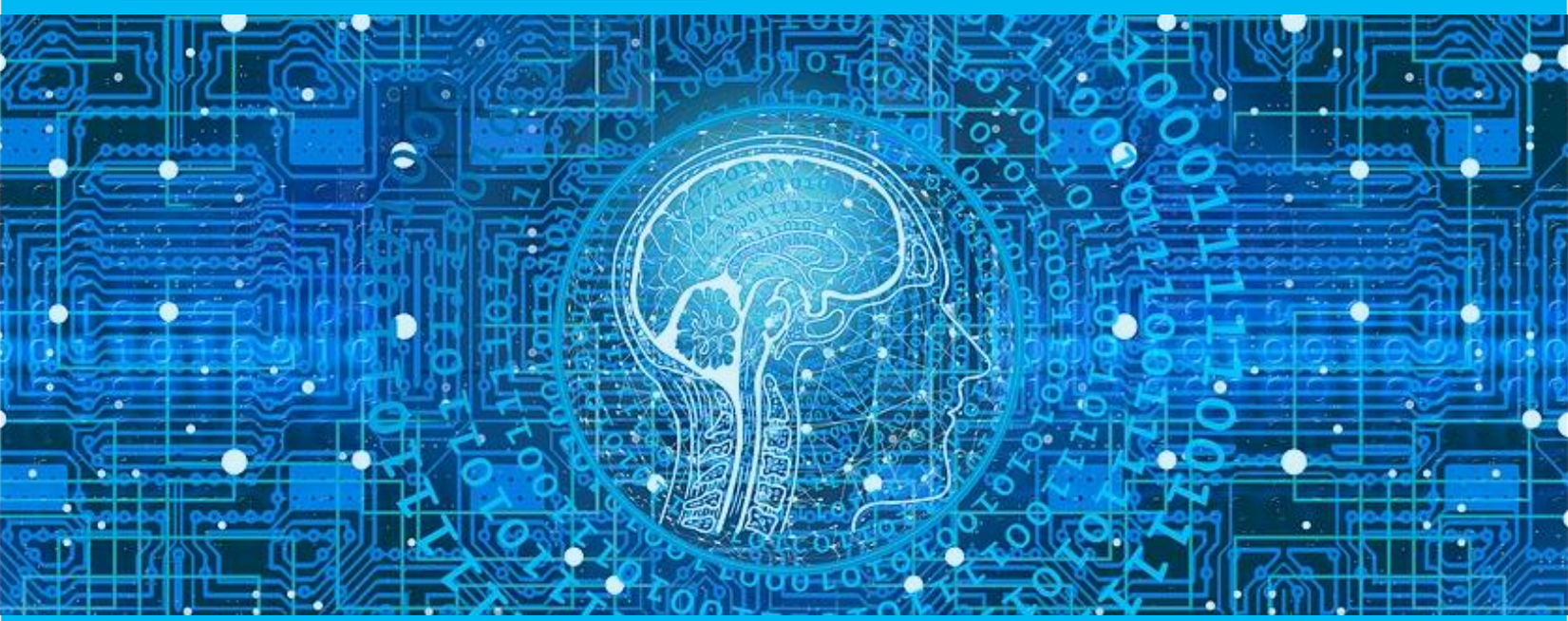
Citizen safety in governmental AI Supported decision-making

An explorative and holistic system's perspective using design science for innovative research



Z.Z. (Zara-Vé) van Tetterode
Engineering & Policy Analysis
TPM, TU Delft

TU Delft

Berenschot

Citizen safety in governmental AI-supported decision-making

An explorative system's perspective using design science for innovative research

by

Z.Z. van Tetterode

to obtain the degree of Master of Science
at the Delft University of Technology.

Student number: 4577701
Project duration: Februari, 2022 – November, 2022
Thesis committee: Dr. H. G. van der Voort, TU Delft
Dr. ir. R. I. J. Dobbe, TU Delft
Msc(Res). R. Postma, Berenschot

TU Delft

Preface

You are reading the research that marks the finalisation of my master's Engineering & Policy Analysis at Delft University of Technology and is conducted in cooperation with Berenschot. To you, the reader, I want to thank you for taking the time and effort to read this part, because the past six years have been an amazing roller coaster, including this report.

During my years in Delft, I discovered many different topics which I could explore during my master thesis. With an interest in many different ideas, especially those with a current societal relevance, researching governmental artificial intelligence-supported decision-making including a direct impact on citizens and how we can keep those citizens safe against government decisions was a perfect fit. I thrive on complexity, finding system's solutions for, in some cases, oblivious challenges, where I can put my creative mind to.

This master piece would not have been born without the feedback, criticism, and inspiring and fun conversations with my graduate committee. In the past years that I have been wandering the halls of my faculty, one person has guided me many times through the jungle of education. Haiko, I want to thank you for all the critical discussions we have had, not only for this final thesis in which we have had many great talks and attempts to play a game, but throughout all of the projects you supported me with your feedback. A special shout-out for my second supervisor as well. Roel, thank you for always asking challenging questions and for the discussions with a philosophical touch. Finally, I want to thank my external advisor, Rosa-May. RM, it has been a pleasure working together. Always critical, with a touch of humour, and including a different practice perspective. A great thanks to you all, I wouldn't have liked research as much without your endless enthusiasm.

A special thanks to my colleagues at Berenshot, including Lars, Lisa, Jill, Mariam, Harro, Freek, and Bart, whom have created an amazing and open learning environment during my internship. Additionally, I want to thank the TU Delft employees whom I have learned so much from. First of all, Nanette, for always having the answer to the question I had not thought of, Peter and Monique, for bringing joy. Rob and Tim for letting me experience the life of an executive decision-maker.

After thanking all of these people, it is left to say that my past years as a student are defined through the the most lovely people in it - corny but correct. I could not have done this without all of my lovely and inspiring friends. Because of them, the past six years have been an amazing journey, where I have developed myself in all ways that the first-year me would not believe in. A big thanks to my parents, Jan & Monique, for always being interested in my daily life and thinking positive. Thank you, Jolijn and Susan, for supporting me in the process of my thesis, but above all for always pulling each other to the next level. Thank you, Yanic, for being my entertainment and introducing me to new things during the endless lock-downs. Regina, thank you for your innovative mindset, support, and obvisously, the coffee. I want to thank my green-yellow friends for the crazy times, the red ones for the high-potential mindset, and the one that always arranges a cake for my birthday for all the laughs. Thanks to my EPA people, to my pandemic friends, to my Barcelona people and mi familia over there. Thanks to everyone who has supported me in the past years in any way. We've made it!

Nos vemos,

Z.Z. van Tetterode
Delft, November 2022

Executive Summary

Governmental AI-supported decision-making is paramount when impacting citizens. Citizens are subject to their government's decision-making, which is crucial when they transact, as examining the transaction's rightfulness is executed by the same government. Thus, control and monitoring are evident, which are increasingly applied through AI-supported tools (Hoekstra et al., 2021). It is deemed socially unacceptable when government falsely accuses their citizens of unrightful transactions. Vice versa, it is also deemed unacceptable when criminals taking advantage of public money are exonerated. The citizens whom are trapped in these systems, like the benefits system, are often vulnerable and have low incomes. The Dutch government wrongly accused 20.000 parents of fraud, resulting in major consequences. Therefore, research is crucial for both a better functioning government and protecting the safety of vulnerable citizens. Scientific relevance is emphasised by the narrow understanding of how harm of next similar case is prevented, as such systems are barely researched holistically. Limited understanding of the relations within and between socio-technological contexts and the safekeeping of citizens lead to the main question:

"How can citizen's safety be safeguarded in governmental AI-supported decision-making?"

Design science is the main methodology, consisting of three cycles: the rigour cycle, relevance cycle, and design cycle (Hevner, 2014). The method allows for innovative thinking and combining empirical (relevance) and scientific (rigour) knowledge toward exploring a new solution, in this research, safeguarding citizen's safety (design). Additionally, it allows for a system's approach required to understand the relationships of different decision-making components and for combining empirical and scientific insights (vom Brocke et al., 2020). Combining this with a holistic perspective, the research incorporates strategical, tactical, and operational challenges and specifically including the political dimension; more than factors and actors. Insight is created between the different context, which are all crucial in establishing a well balanced system.

In seeking an answer to the main question, three parts are crucial: *the system* of AI-supported decision-making, the concept of *citizen's safety* and how both can be integrated. *The system* and its boundaries are discovered through scientific and empirical exploration, serving as a framework. The definition of *citizen's safety* and its implications are discovered by relating to the system and its characteristics. With these two parts, the influence of *the system* on *citizen's safety* is explored and leads to the third crucial part, validated by a serious game. All parts use different sub-methodologies to obtain the required knowledge for the design science methodology.

The system is defined by its decisions in the political, organisational, or technological context, conducted through actor and system analysis (Enserink et al., 2010), based on synthesising the scientific integrative and semi-structured literature review (rigour) (Snyder, 2019), the policy and law review in the European Union and the Netherlands (relevance), the case study of the benefits system (relevance) (Johansson, 2007), and expert interviews for validation (design), resulting in *the system*, its boundaries, and its properties through all three cycles. The crucial properties include the contexts in which decisions are made, where *the system* behaves in a societal context. In the open *system*, society shapes public debate, resulting into laws formed by politics in *the system*, resulting in policy formed by organisation, the execution of said policy, and the process of creating a suited algorithm. The final decision is made in the organisational layer, minding the algorithm's outcome, by the final human decision-maker, and affect the citizens subject to the system. Two inner system loops exist, (1) among the citizens that are subject to *the system*, being researched every specified time *research loop* and (2) the training data used to enable the algorithm to determine the weight of the predictive variables, the *data loop*. Furthermore, the multi-actor and multi-disciplinary characteristics are crucial, resulting in opportunities for actor conflict and uncertainties, adding to complexity through information asymmetry. The outcome of *the system* is unknown in real-time, if ever, as it is timely to prove one is guilty, and vice versa, proving one is innocent is unattainable. *The system* knows two objectives which behave as a sinus over time, consisting of detecting fraud and providing service. These objectives are dependent on public

and political opinion, resulting in difficulty of defining safety in an absolute way. Combining the system characteristics lead to the conclusion that the problem operated in *the system* is inherently wicked.

Citizen's safety is specifically referred to in the context and characteristics of *the system* and ought to cover safety of citizens within *the system*. Due to the wicked nature of the system's decisions, interventions are not straightforward. Thus, safety measures, as defined in literature (Dobbe, 2022; N. G. Leveson, 2011), cannot be applied directly, especially concerning the political and organisational context (rigour). Due to the unknown outcome, preventing false negatives and positives is an unlikely approach. However, also under changing system objectives and lacking vertical integration, citizens should be kept safe by and from their government. Therefore, *citizen's safety* is defined to aid in preventing harm to citizens by combining the scientific and empirical literature regarding AI systems as a good balance between equality, privacy, and transparency. A guaranteed value-balance in *the system* can potentially prevent the harm done in the childcare benefits case, as technological decisions were based on citizens having second nationality, low income, and single-parent families, and politics emphasised foreigners as fraudsters, trickling down to organisations (relevance). Additionally, these values have a legal base in the Netherlands and the European Union (relevance).

Combining both concepts in a conceptual systems diagram (design) brings insight into the most influenced components by plotting them on each other, which are the training data and the algorithm training. As these components lay in the *data loop*, the loop becomes more influential. *citizen's safety* cannot be calculated due to unknown population values yet can be connected to the system objectives, resulting in a conceptual systems dynamics model (design). Additionally, the influence on the objectives plays a central role as well. It is deduced that when striving for equality, transparency, and privacy, servicing citizens becomes a more central notion. Vice versa, for fraud detection, the influence lies in decreasing transparency and privacy, which is an expected result. The relationships among the values are validated by a semi-empirical serious game (design). This game additionally discovers that the value's preference shifts depending on the actor. The final decision-maker is played by the respondent, following the human-on-the-loop principle inherent to *the system*. The shift in value per actor can determine that *citizen's safety* is viewed differently depending on the actor, adding complexity to designing interventions. Combining this insight with the most influential components laying in the technological context, the final decision-maker is not able to influence *the system* for a good *citizen's safety* balance, even though they are relied on for the human-on-the-loop principle, where they ought to function as the safe-keeper of *the system.*

Concluding, *citizen's safety* decisions must explicitly be made before the *data loop* starts by the political and first organisational decision-makers, as it is influential. However, it is empirically unrealistic to assume that all decisions can be made preceding operation of the loop. Therefore, the actors in the *data loop* require feedback means in which timing is crucial. Secondly, the final decision-maker lacks the means to prevent or execute decisions when they disagree with the algorithm's outcome, even though they are relied upon through the governance structure and the loops end with this last decision for citizens found guilty. Thirdly, the differences between human-decision making occurring at random and non-human decision-making occurring systematically are not acknowledged in the government's stance on and the execution of AI-supported decision-making. As there is complexity in predicting whether one is receiving a transaction rightfully, mistakes must be handled properly, which results in the need for responsibility-taking, which is not directly present in the childcare benefits case. Human mistakes are often made at random, thus disadvantaging all groups equally over time. Algorithmic decision-making disadvantages specific groups. This bias also roots in the actors present in *the system* and specifically in the loops. This crucial characteristic of AI-supported decision-making must be acknowledged by everyone who decides in *the system*.

Recommendations include research into the role of monitoring and control from the inspectorates and the judicial system in *citizen's safety*, but only if they act timely. Additionally, a definition of well-balanced *citizen's safety* in the context of wicked problems requires both academics and practitioners to contribute. This research takes the first steps, but with collaboration with practitioners and academics the safety notion and its thresholds can be calculated. Thereafter, intervention and scenario analysis can be conducted to critically assess the effect on *citizen's safety.* Lastly, the recommendation includes expanding the serious game to cover a multi-player learning environment for the system's decision-makers to gain knowledge on how their decisions affect *citizen's safety.*

It is deemed unacceptable for the next childcare benefits affair to happen, and this research helps to create a holistic understanding of its system.

Acronyms

AI Artificial Intelligence

ADM Automated Decision-Making

BCO Management Support; Governance & Support

CCB Case childcare benefits

EU European Union

HIC Human-in-Command

HITL Human-in-the-Loop

HOTL Human-on-the-Loop

KPI Key Performance Indicator

Ministry of SAE Ministry of Social Affairs & Employment

SAB Social Assistance Benefit

WI Work & Income

WRR Dutch Scientific Council

Glossary

Actor refers to a stakeholder, an organisation or other entity with a concern in the problem.

Childcare Benefits Case refers to the 'Toeslagenaffaire' in Dutch. In this case, many parents are wrongly accused of fraud and therefore helps explains how unsafe cases work, and what such a system looks like.

Design Science refers to the main methodology of this research and mainly consists of three cycles: the rigour cycle, relevance cycle, and design cycle (Hevner, 2014).

GDPR refers to the General Data Protection Regulation (GDPR) (2016), taking into account the privacy of European Union's citizens.

Governmental AI-supported decision-making refers to governmental decision-makers using a technological component, like predictive modelling, in the form of artificial intelligence to substantiate their decisions. The decision is eventually decided by a human decision-maker, which refers to the term supported. Also referred to as "the topic".

Social Assistance Benefits Case refers to the social assistance benefits case, where an algorithm is used for predicting fraud, similar to the childcare benefits case. In this case, the operational context is explained in-depth.

System Objectives refer to the objectives of The System, including providing service to citizens and detecting fraud.

The System refers to the designed system in chapter five, where the scientific and empiric explorations are synthesised to a system consisting of decisions in the contexts of the socio-technical system.

List of Figures

Contents

01 Introduction

Content

1 Introduction

Governmental organisations are increasingly impacting citizens' lives with Artificial Intelligence (AI)-supported decision-making. The outcome of such decisions may aid or harm citizens, as the decisions may result in safe or unsafe situations for citizens. The decision's reach may be unaccounted for by the decision-makers preceding the decision's execution. The may haves refer to an actual case, the Childcare Benefits Case, or *Toeslagenaffaire*, in which governmental decision-makers, with the help of AI determined profiles, falsely accused 20.000 parents of fraud, leading to massive debts, out-of-home placements of children, and even suicide (Amnesty International, 2021; Frederik, 2021b; Grol, 2022; Henley, 2021a). Ultimately, it turned out most behaviour is considered to be innocent today. How such harm is averted in future cases is yet unclear.

This thesis adds to the finalisation of the Master of Science in Engineering & Policy Analysis at the Delft University of Technology and the digital innovation in public organisation's internship at Berenschot. This chapter introduces the research, starting with the problem specification in 1.1, followed by the research design in 1.2. The section 1.3 argues for the scientific and societal relevance. The relevance for the masters programme is explicitly illustrated in section 1.4. This chapter concludes with the thesis outline, including the research flow diagram in 1.5.

1.1. Problem specification

This section aims to clarify the problem, which is the impetus of this research. Starting with the problem context in section 1.1.1, elaborating on the frame of reference for the problem, after which the knowledge gap and aim are explored in section 1.1.2. The latter entails the main research question as well.

1.1.1. Problem context

The problem context refers to the context of the decisions made by government, supported by AI and is explained in this section. *Governmental AI-supported decision-making*, as posed in the introductory text of this chapter, can lead to a grove impact on society. The term refers to governmental decision-makers using a technological component, like predictive modelling, in the form of Artificial Intelligence (AI), to substantiate their decisions. The decision is eventually decided by a human decision-maker, which refers to the included term supported.

One poignant example is a case that has stroked much public attention since 2020: the Childcare Benefits Case. Much debate has followed regarding this topic, both in society, politics and organisations. In this case, 20.000 parents were wrongly accused of fraud, resulting in massive debt, out-of-home placement of their children and even suicide (Amnesty International, 2021; Grol, 2022; Henley, 2021a). Since the case came to light, numerous arrangements have been executed by the *Tax Authority*, the stakeholder responsible for the collection of benefits (Tax Authority, 2020), however, the heavily affected citizens remain unaided (Leendertse, 2022). The case only came to light after the harm was done, and the citizens affected by the government decisions were not believed for a long time.

The system of the Childcare Benefits Case is judged to be discriminatory by the Autoriteit Persoonsgegevens (2018) as the data included the second nationality of citizens. Notwithstanding, more was to it than the inclusion of the second nationality in the technological application. Citizens were not trusted when they submitted an appeal against the decisions made about them—neither by the executing agency, the Tax Authority, nor the judge. After attracting massive media attention, financial compensation measures are implemented for the harm they endured. However, the application process is bureaucratic, and not all children are reunited with their parents, some of whom have been having problems due to fraud accusations since 2013, hence, for almost ten years (Frederik, 2021a, 2021b). Ideally, the negative impact on citizens, including vulnerable children, was avoided.

The Dutch government is increasingly using AI-solutions, with 165 active applications (Hoekstra et al., 2021). These applications have discrepancies in pursuits. The one to most common objective is situated within the field of inspection and enforcement: detecting crime. This is a politically and societally sensitive subject, as opinions differ on the origin and the predictors of crime and one is deciding on citizens' lives. How the government handles safeguarding citizens precisely is opaque, as either publicly available information or their knowledge about this topic is lacking. What is known is that the *Childcare Benefits Case* used a self-learning algorithm trained to recognise fraud from the population of childcare benefits receivers. The algorithm determined the variables required to predict and assigned their weight. This type of algorithm is part of the monitoring and enforcement category, which is the most popular domain for AI applications.

Knowledge about AI-supported decision-making in socio-technical systems is lacking scientifically and empirically. Exquisitely, what should be done to avoid immense impairment to citizens is unknown. One reason for the severity of this challenge is that citizens cannot choose a different government and will invariably be the entity of influence, whether desired or not. It is intricate for the government to make a suitable trade-off between preventing criminality and providing service to citizens. To manually investigate the population of benefits receivers for fraud comes with capacity problems, therefore technological support is implemented to point in the right direction. However, pointing in the right direction through technology may lead to bias, privacy breaches, and lacking transparency as a side effect (Agbozo & Asamoah, 2019; Cerquitelli et al., 2017; Marda, 2018). One probable resolution can be to execute safety measures in the system. However, how safety relates to the system described by *Governmental AI-supported decision-making* is not researched. In short, complexities, uncertainties, and conflicting opinions exist regarding this topic.

1.1.2. Problem statement & knowledge gaps

This section conveys the knowledge gaps on *Governmental AI-supported decision-making* to usefully position and define the problem statement and main research question. Due to the iterative nature, feedforward references are made toward the body of this research. The scientific knowledge gaps are defined through the scientific exploration, chapter 3, through the exploration and connection of the notions Artificial Intelligence (AI), wicked problems, socio-technical systems, and safety. The societal knowledge gaps are defined through an empirical exploration, chapter 4, where policy is analysed and the benefits system is explored.

The overarching scientific knowledge gap is the lacking knowledge on the system of governmental AI-supported decision-making. The scientific exploration in chapter 3 shows the evident knowledge gaps, which are discovered through an elaborate literature review focusing on the notions AI, wickedness, sociotechnical systems, and safety. The underlying connection between the notions is the governance structures that shape them in a governmental context. The scientific knowledge gaps are:

- *AI challenges from a holistic perspective*
  AI challenges are regarded from individual perspectives, yet lack a holistic approach. The challenges include intrusion of privacy, bias (equality), and transparency. A holistic approach can add the interconnections between the challenges to better understand the trade-offs that may occur in the system and take into account the both a policy and technical perspective on the challenges.

- *Collision or connection between political governance and operational safety*
  The governance perspective on the political and organisational level is mostly strategic, while the notion of system safety entails a very operational approach to the sociotechnical system. It is

unknown how both concepts can connect or collide for the system of governmental AI-supported decision-making.

- *Operational wicked problem behaviour*
  Wickedness is researched in the context of policy problems, not how a wicked problem behaves in an operational context. Even though wickedness is a contested notion in science, it does help to identify the problems at hand. Agreement can be found in the characteristics complexity, conflict, and uncertainty. How this influences the operational context is unknown.

- *Feedback structures in governmental AI-supported decision-making*
  Ideally, a real-time feedback structure on the quality of decisions is present in the system. Currently, the feedback system in the sociotechnical system is understood for the citizens that are affected by the decisions and their concerned social circle to shape public opinion and influence the system with their democratic rights. It is unknown how the vertical integration is structured within the system of governmental AI-supported decision-making.

The overarching societal knowledge gap is the lacking knowledge on how to protect citizens, while detecting criminals in high-risk governmental systems. Criminality detection systems benefit society, yet can disadvantage citizens as well. The balance is unknown, and one right answer cannot be argued for indefinitely. More concretely, the void lies in the knowledge about decisions' influence throughout the complete decision-making process on contributing to the conclusive safety outcome of said system.

- *Impact of Human-on-the-Loop (HOTL) governance structure*
  The HOTL governance structure is observed in the empirical exploration and the impact of human oversight is contested in the scientific exploration. In what manner the governance structure impacts the system and affects citizen's safety is empirically not observed, except the harm detected in the Childcare Benefits Case and the harm not detected in the Social Assistance Benefits Case.

- *Citizen's safety in the system*
  The Childcare Benefits Case shows the trade-offs that are made by the political actors involved between detecting criminality and providing a service to citizens. The direction in which is steered to keep citizens safe is elaborated on empirically, however, is not elaborately illustrated in both cases.

- *Values as a safety net*
  As the political context is part of the system, laws and regulations become dynamic. Therefore, another way of executing safety is required than the usual law - policy - measures cycle in safety. Above the national laws are international and EU laws which often come down to striving for certain values if noted for protection against government. The question arises if values can help create citizen's safety and how they are present in the system.

- *The multi-actor environment*
  The environment of the system regarding governmental AI-supported decision-making is prone to many governmental actors, all who have their own rights, tasks, and behaviour. The actors involved play an extensive role in the decision-making processes and, therefore, more knowledge on the actor arena is desired.

This research aims to attain knowledge on safeguarding citizen's safety in a governmental decision-making system using AI in an operational context. Hence, the main question is defined as follows:

"How can citizen's safety be safeguarded in governmental AI-supported decision-making?"

The question is further elaborated on in chapter 2, in which the sub-questions are defined. Preceding that chapter, the research design is presented in the next section by expanding on the scope, approach, and main methodology.

1.2. Research design

This section defines the research design that answers the main question. The scientific and empiric explorations are conducted respectively in chapters 3 and 4 help define the research design, elaborated on in chapter 2. Through this iterative process, the research is designed. This section starts with the approach and methodology is elaborated on in 1.2.1, followed by the scoping decisions in section 1.2.2. Lastly, the specific boundary choices are explained in the delimitation section 1.2.3.

1.2.1. Holistic approach to design science

The overarching method of this research is design science, combined with a holistic approach, and a system's perspective. The combination of these results in an insightful exploration of design artefacts (design science), including The System (system's perspective), including the capture of wide range of challenges (holistic approach).

Design science requires iterative interaction between different research stages to answer the question meaningfully. The advantage is the allowance for creative implementation of the approach to this research, as design science is about synthesising knowledge from different types of sources toward a design. The disadvantage is that it can be challenging to follow the chronological order of research due to its iterative nature. Therefore, every chapter elaborates on the research phase, depicting where the chapter stands concerning design science and its goals. Additionally, different parts of this research explain how certain elements are concluded. As the elements are needed to explain other parts purposefully and are not always in sequence, the text entails references feed-forward.

The holistic approach results in the inclusion of challenges that may be overlooked when focusing on either the human or technological entities. Additionally, it allows for a broader perspective of decision-making, not just resulting from the political dimension that is included, but also allowing space for the challenges of decision-making captured by the wicked characteristics of the problem affected by the decisions in The System.

The system's perspective is catched not only by actors and their decision on actors, but more broad also by the wickedness of The System's problem and its safety. The system's perspective allows for insight in interconnections between distinct components, not captured before in this context. Because of this perspective, it is a challenge to define what is good and what is bad for safety, as the political dimension is included and solutions to wicked problems are never 'right' or 'wrong' (Rittel & Webber, 1973).

The approach is exploratory for an in-depth understanding of the problem, which suits as the scientific background of this topic is meagre. Even though literature is lacking, Governmental AI-supported decision-making systems are empirically implemented. Iterations are inevitable in an exploratory environment, which fits the iterative character of Design Science.

A mixed methods approach comes to the meaningful details for the design cycles. A literature review is used first in the rigour cycle, while the relevance cycle is explored empirically through desk research and a case study on benefits. After this, the rigour and relevance cycle start the design cycle through a systems analysis and a conceptual systems diagram model. Lastly, an artefact is designed to validate, thus going back from the design cycle to the rigour cycle. The cycle is completed by the conclusion and recommendations, flowing back toward both the empiric and scientific world.

1.2.2. Scope

Scoping decisions are made to focus on a specific problem. The scoping decisions made for this research are gathered through the scientific exploration in chapter 3 and the empirical exploration in chapter 4. The scoping decisions are:

- *Direct impact on citizens by the decision*
  Firstly, *Governmental AI-supported decision-making* is scoped to cover the decision-making that has a direct impact on citizens, as this type of decision-making is most urgent to keep safe. Therefore, the research is scoped to cover those AI algorithms that cover impact on citizens, often using personal information of citizens to come to a conclusion. This scoping decision results in taking certain values in account, as those are substantiated by laws. For example, when dealing with personal information, the processor of that information needs to be compliant with the General Data Protection Regulation (GDPR) (2016).

- *The Netherlands as geographical focus area*
  Secondly, the geographical boundary of this research reaches the Netherlands, as the Netherlands is front runner in AI applications for decision-making (Koens & Vennekes, 2021), and upholds democratic rights to citizens (Constitution, 2018). The results of this research may be applied to similar countries and the European Union as well, although special caution needs to be taken around the organisational context, as assumptions and results are defined for the geographical scope of the Netherlands, and every state apparatus or polity may be different.

- *Wicked problem's decisions*
  Thirdly, several notions are placed explicitly within the scope of this research, including AI, socio-technical systems, safety, and wickedness. The assumptions, results, conclusions, and other statements require to be taken account in this context. What the notions mean and the relation among them is explored in depth in chapter 3. For the scope, wickedness means that conflict, complexity, and uncertainty are present in the decision's system through the problem that is handled. This results in more complex decision-making measurements to evaluate the decisions and the system. As societal problems are often wicked, this scoping decisions also functions as a system characteristic.

- *Human decision-making with supportive AI*
  Fourthly, this research focuses on AI-supported decision-making, and therefore Automated Decision-Making (ADM) is out of scope for this research. ADM refers to the final decision originating from a technological component, a non-human decision-maker. This research explicitly states that the final decision-maker in the system is human, therefore is scoped to AI-supported decision-making. This is of importance as different rules apply for ADM than for AI-supported decision-making regarding impacting citizens and using their personal information (General Data Protection Regulation (GDPR), 2016).

1.2.3. Delimitation

This section defines the boundaries of this research. Throughout the research delimitation decisions are made, as this research concerns an iterative character, specifically through the scientific exploration in chapter 3 and the empirical exploration in chapter 4. The delimitation includes:

- *Interaction between inspectorates and their governing body, i.e. ministries*
  Firstly, four contexts are differentiated for this research, one of which is the organisational context. The organisational context includes the involved actors defined in 5, to which demarcation decisions are constructed. The interaction effect between inspectorates and their ministries is left out of the scope, as the inspectorate on national benefits is only construed since 2022 (Official Gazette 2022-4749, 2022), and evaluation is mostly conducted by the independent Council of Audit, regarding public spending (Compatibility law 2016, 2016; Constitution, 2018). Several inspectorates conducted research after the wrongdoings in the Childcare Benefits Case became apparent (Inspectie Justitie en Veiligheid, 2022; Inspectie Overheidsinformatie en Erfgoed, 2021), yet how they meddled during the time citizens were wrongfully accused of fraud is unclear. Therefore, inspectorates are deliberately excluded in paving a way to a safe system.

- *Judiciary system is not the primary factor*
  Secondly, the feedback loop supported by the judiciary system is not suited for direct feedback toward changes in the system, and therefore is not the prime factor of research. Argumentation includes that (1) the final ruling takes too much time for direct feedback, (2) the harm is presumably already conducted, and (3) the court follows the law, seen in the Childcare Benefits Case, in chapter 4. Therefore the solution for a safe system is to seek insight into the system before harm gets tangible.

- *Subcontractors developing algorithms*
  Thirdly, the subcontractors used by governmental organisations to develop the AI are out of bound. They may be involved in developing, yet they must always behave under the jurisdiction given by the problem owner, the governmental organisation. It is assumed that the rules and regulations can be enforced if the government has the proper knowledge and testing capabilities. Additionally, the government is also accountable for its hiring.

- *Intervention analysis*
  Lastly, interventions on the system for enhancing safety are out of scope, as this research intends to explore, not to find one solution. The exploratory nature is argued in the next section. The solution requires dedication and engagement from both governmental and scientific researchers involved, as more specific research is needed to be able to test solutions. The specific research should then add to this holistic approach.

1.3. Scientific & societal relevance

This section explains the contributions and relevance for both science and society. Both relevance types are paramount, as it gives value to the research.

The central scientific contribution is the multidisciplinary systems approach towards safeguarding governmental AI-supported decision-making systems, which can be applied to all fields where the outcome directly impacts citizens decided on by humans and non-humans. The unique relevance of this research for the scientific community combines a new emerging fast-pacing field, AI-supported decision-making, in a matured institutional context of governments that have existed for centuries where change often means disruption. Additionally, the value of this research is added through the combination of the empiric field and scientific substantiation through an innovative design science methodology, creating the ability to add to the knowledge base in a meaningful way.

The primary societal contribution is gaining knowledge on preventing harm to citizens, bringing academic and empiric characteristics together through analysing cases and conducting a semi-empiric experiment on the effect of governmental AI-supported decision-making. Citizens play a central role in this research, as the wicked characteristics, undefined citizen's safety, and lack of individual power contribute to a difficult position for them.

In more practical terms, the societal relevance is empathised through the relation between citizens and government. When government is expected to make decisions, as is the case for societal problems, e.g. fraud in public money, citizens need to be protected by that same government. For this reason, the independent judicial system and Ombudsman are constructed by the constitution (Constitution, 2018). However, through the Childcare Benefits Case it is proven that these institutions cannot always protect citizens before harm becomes extensive. As scientists and practitioners do not precisely know the causation leading to this failure, it is evident to conduct an empirical analysis. With the societal aim to research how such harm can be prevented, this research knows a high societal relevance.

This research challenges academia and practitioners as it not only explores the system but, additionally, the first stone is laid in the innovation of jargon for safety in governmental AI-supported decision-making in a wicked environment.

1.4. Link to the master Engineering & Policy Analysis

This research is related to the master's Engineering & Policy Analysis of Delft University of Technology in several ways. The master's program is aimed to systematically analyse grand challenges from multiple disciplines, complemented by modelling or simulation techniques. The finalisation of this program is captured in this research and excellently links to it.

Grand challenges relate to a challenge with a technical and political component, e.g. climate change, poverty, and pandemics. The grand challenge central in this research is the safekeeping of citizens in Artificial Intelligence (AI)-supported decision-making, which directly relates to the sixteenth sustainable development goal of the United Nations (n.d.) regarding strong institutions, including accountable and transparent organisations, and equal and privacy protected decision-making. The technological component relevant for the system is the solution for the detection of criminals and organisational capacity problem: the AI application. The politically relevant component can best be illustrated over time in the Childcare Benefits Case, relating to contrary political goals on how to provide service to citizens, further elaborated on in the next chapters.

Additionally, this research is analytical through the methodologies used, analysing the synthesised explorations to demarcate The System. Complexity in this topic is found through the system's perspective and multi-actor characteristic. Even when all actors are governmental, therefore having complementing goals, the actor arena knows challenges, e.g. information asymmetry may result in actor conflict. Methods characterising the master's program are used, as the system and actor analysis (Enserink et al., 2010), conceptual systems dynamics model (Keys, 1990), and a serious game (Van

Daalen et al., 2014; Van Der Zee et al., 2012).

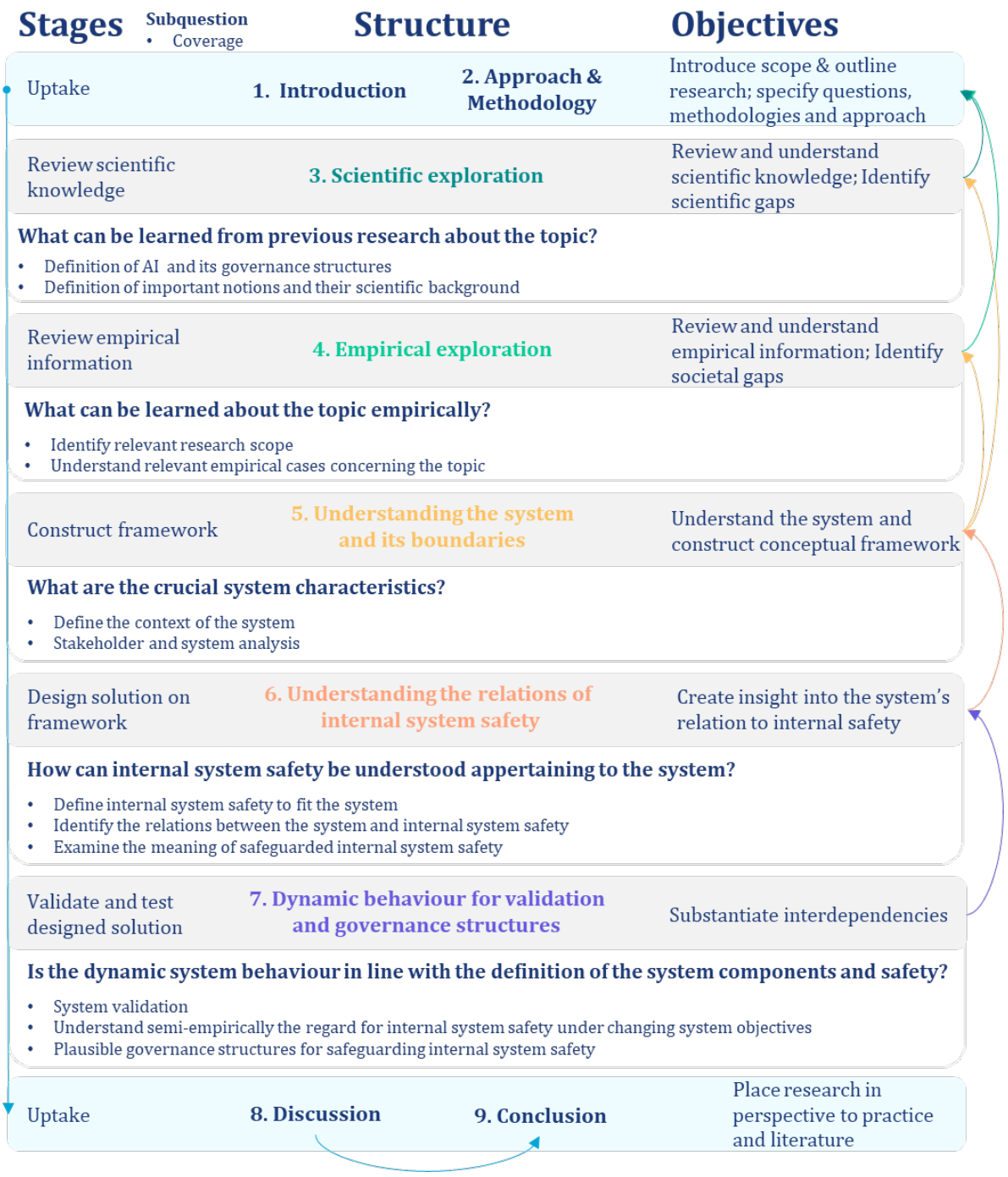
Lastly, through the benefits study the research becomes tangible, often a challenge when dealing with grand societal challenges. The wickedness presented in this research has not been researched in the operational context of AI-supported decision-making, which is new. As the master's program equips its students with multi-disciplinary methods to explore undefined and uncertain challenges, this research was possible.

1.5. Structure of research

The goal of this section is to give oversight to the outline of this research. In the next chapter, chapter 2, the approach and methodologies used in this research are argued for in detail. The foundation of the research is laid to come to a meaningful way of answering the main research question. Here, the sub-questions and accompanying methodologies are explained as well. Next, the scientific exploration is presented in chapter 3. The scientific consensus on different notions is portrayed, concluding with the major scientific knowledge gaps in current literature. After that, the empirical exploration is presented in chapter 4. The policy and cases depicted cast an important role in understanding the complex decision-making context, which is brought together in chapter 5. Concluding this chapter are the system characteristics and its boundaries, presented both textual and graphical. Chapter 6 continues with the presented system to build on safety, defined in the sixth chapter. Next, the system and its interdependencies are tested for safety to understand better the dynamic system behaviour of values leading to increased citizen's safety in chapter 7. The semi-empirical results contribute to the validation of this research. The discussion and conclusion are refined respectively in chapters 8 and 9 to place this research in its scientific and empirical contexts.

This chapter concludes with the research flow diagram, illustrated in figure 1.1. This diagram depicts the chapters in the middle, including their topic. The left side refers to the research stages. These are defined earlier; however, they also include the uptake, which is significant for research. The objectives are illustrated on the right side, including the objectives related to the chapter depicted in the middle. The chapters relate one-on-one to the sub-questions depicted after the grey bar. Underneath the sub-question, a concise overview of the coverage is inserted. On the right of the illustration, the iterative process is portrayed through the arrows going up again. Especially in designing the system, iterations were significant to come to a useful synthesising of the explorations.

Figure 1.1: Research flow diagram



02 Approach & methodology

Content

2

Approach & Methodology

An apparent oversight of this research's approach and methodology is outlined in this chapter. The iterative character of the main method, design science, results in a discontinuous order of outcomes reshaped logically in sequential order in the following chapters. Therefore, this chapter is crucial in understanding this research. Governmental AI-supported decision-making is the central topic and for societal relevancy recommendations for practitioners are valuable, while for scientific relevance the academic value is crucial. To add value to both science and society, design science is the overarching method for this research, as an innovative and creative mindset are needed, combined with an iterative approach. The main methodology is elaborated in section 2.1. The approach to this research is a sociotechnical system's approach and adds value as all dimensions are taken into account through the decision-chain, from strategical to operational decisions. This choice also requires multidisciplinary knowledge and it is elaborated in section 2.2. Approaching design science sociotechnically, the question arises how citizens' safety is actually safeguarded in the sociotechnical system. How this question is answered is explained in section 2.3, where the sub-questions add to answering the main research question and the methodologies per sub-question is explained. The research assembly shows how the sociotechnical approach and design science are combined to come to insightful conclusions and finally, the added value to both science and society.

2.1. Design science as overarching method

This section reasons the choice for design science and what the methodology entails. The leading argument for design science is that this research entails two crucial characteristics: an elaborate empirical background and lacking scientific substantiations. In design science, both empirical and scientific components play a crucial role in designing, illustrated in figure 2.1.

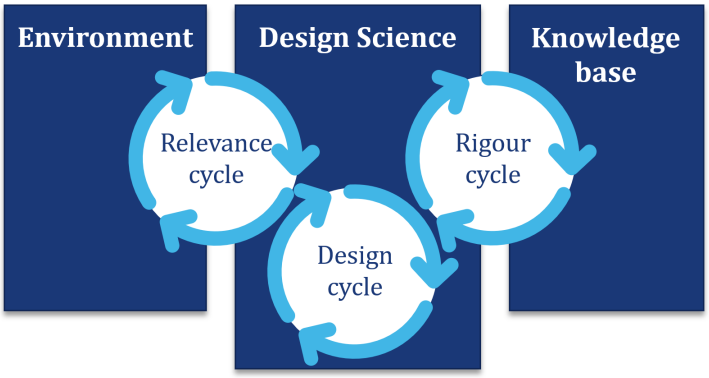


Figure 2.1: Design science cycles, inspired by Hevner (2014, p. 88)

Design is to be interpreted broadly, as the background for this research lies in modelling and sim-

ulating, in combination with policy analysis, which can be viewed as designs. Adding to this argument is that models are created through scientifically and empirically synthesising, after which the research ought to come to a meaningful conclusion, giving back relevant knowledge to both worlds, which is design science by definition.

Design science combines practice (the relevance cycle) and theory (the rigour cycle) toward a solution-oriented design. It is a problem-solving paradigm for which the goal is to gain knowledge of system components and their relations, to make a fitting design (vom Brocke et al., 2020), and is pragmatic (Hevner, 2014). As defined by (Hevner, 2014), three research cycles in design science are recognised among academics vom Brocke et al. (2020), and is depicted in figure 2.1. The relevance cycle starts with the opportunities and problems seen in practice and the criteria to which a solution should hold. The environment consists of people, organisational and technical systems, recognised in this research by defining the societal, political, organisational, and technological contexts, further elaborated on in chapter 5. The design cycle is about designing a solution, or artefact, that can help resolve the problem at hand. Central is the building of artefacts and processes and, thereafter, evaluating them. The rigour cycle focuses on the artefact's evaluation and includes experience-based knowledge and scientific theories. Through the artefact, knowledge is added to this base. Interesting to note is that Hevner (2014) refer to the potential harm when leaning too much toward explaining all outcomes in a scientific and theory-backed manner, as design science allows for innovative ideas.

One disadvantage of design science is the iterative process. Although it suits the problems and works toward a solution, it is undefined when the problem and the artefact are optimised; thus when iterations will not add to the functioning. Additionally, the boundaries of design science are broad and not predefined. Thus the researcher needs to be cautious, especially when the problem itself is partially undefined, i.e. wicked problems. This research explores the general system demarcation and how the solution space can be sought.

As this research regards topics that lack a thorough scientific substantiation yet are already in place in practice, the need for research is evident. Despite the differences between theory and practice shown in the two upcoming chapters, the common objective to come to a better understanding of governmental AI-supported decision-making, how the interdependencies are shaped, and how the system might be evaluated are clear. The rigour and relevance cycle are elaborated upon to work towards these objectives, respectively, in chapters 3 and 4. As these chapters are both meaningful to synthesise for the design cycle, they are also useful, which comes together in chapter 5. The design cycle is repeated for citizen's safety in chapter 6. Lastly, a semi-empirical experiment is conducted to add to the design cycle through the experiment design, the rigour cycle through the validation of the conclusion of the preceding chapter, and the relevance cycle through the experiment results.

2.2. Sociotechnical systems approach

This section aims to reason the choice for a sociotechnical systems approach, explain what the approach entails, including its scientific background, and how it is reflected in the coming chapters. The main argument for the systems approach is that interdependencies exist between the different parts of governmental AI-supported decision-making. They are implemented empirically without being researched scientifically. The reviewed system is sociotechnical by nature, including both technical and social disciplines, which means that evaluating such a system requires fitting both disciplines.

More specifically, many researchers have deep-dived into specifics of the system (OECD, 2019), like threats and opportunities when using technology, how to design functioning organisations, and how to embed public values into laws. However, how the aspects combine into one system, including a social and technological perspective, is often overlooked. The knowledge gap between making policy and implementing technology is considerable, and even though knowledge is lacking, the system is operated. Baxter and Sommerville (2011) also argue that a sociotechnical approach is often lacking, therefore missing the realisation of the potential benefits. Bringing the interdependent social and technical components together for understanding and improving is the essence of the sociotechnical systems theory approach (G. H. Walker et al., 2008). An example of the sociotechnical systems approach related to this research topic is trust in intelligent systems (Benk et al., 2022; Jones et al., 2013), where Benk et al. (2022) even goes as far as to attempt measuring the trust in the system. However, concluding that trust in a sociotechnical system might be too complex to measure. This example illustrates the advantage of a systems approach when the goal is to be relevant empirically and

scientifically. A systems approach in this research means having a holistic perspective, as the focus is on the interdependencies and relations among system components (Wu et al., 2015).

The disadvantages of this approach are that experts on system components have more knowledge about the functioning of the component and how to influence their part of the decision-making process. Politicians are experts in making laws, ministers know specifically about their field, and executive agencies know how to implement the decision-makers' will. Even though these advantages of a specific expert's approach are lacking, the advantages of a systems approach are deciding. In conclusion, the sociotechnical systems approach is decided, primarily as this research focuses on the interdependencies between system components and an overall approach to safety.

In the coming chapters, the sociotechnical approach firstly is seen through the elaborate scientific background in chapter 3, after which the social aspect of the systems perspective in both policy, organisation and law are present in chapter 4. Chapter 5 designs the first general system, including technological, organisational, political, judicial, and societal perspectives in which the systems perspective is evident. After that, the system is used to interpret safety in the different disciplines in 6. Lastly, the systems approach is again depicted in the experiment in chapter 7 as the research uses both qualitative and quantitative data to come to conclusions, and the strength of the relationships is evaluated.

2.3. Research assembly

This section gives insight into the sub-questions and how they are answered. Starting with the overview and order of the research questions in section 2.3.1, followed by section **??** elaborating on the methods used for that sub-question. This section concludes with an overview of the research stages, depicting the relation to design science as overarching methodology and the methodologies per sub-question in section 2.3.2. This section gives oversight to this research and explains how the research is conducted.

2.3.1. Research questions & methodologies

This section amplifies the sub-questions and their accompanying objectives to illustrate that the answers to the sub-questions form the main research question's answer. Additionally, the methodology is elaborated on in this section. This section starts with the main research question, followed by the sub-questions. The main research question is defined as follows:

How can citizen's safety be safeguarded in governmental AI-supported decision-making?

Governmental AI-supported decision-making refers to the system of governmental decisions that lead to a final decision which influences citizens. The decisions in this system are all made by governmental bodies. The term "AI" is added to underline the self-learning and complex technological context present in this system. The term "supported" is added to the form in which the technological context aids the decision-making process because ultimately, a human makes the final decision, and the technology is an instrument. For the substantiating argumentation, refer to chapter 5, in which the system is designed and illustrated.

Citizen's safety refers to safeguarding the internal system from itself and towards citizens. In this, the unintentional harm that can be done to citizens is captured. This effect can be taken into account in this terminology because the governmental actors included in the system aim to take care of citizens and keep them safe. Thus, in general, citizen's safety refers to the decisions made in the decision-chain to keep citizens in the system safe. Refer to chapter 6, in which citizen's safety is defined and is related to the system objectives.

Five sub-questions support the main research question. Understanding the system through previous academics and evaluating what they examined is the first step in conducting this research. It results in meaningful definitions of AI, its governance structures and the scientific background on the essential notions, leading to the first sub-question:

1. What can be learned from previous research about safeguarding governmental AI-supported decision-making?

In answering this question, scientific knowledge gaps are discovered.

The first question is answered through a literature review focusing on peer-reviewed academic literature with a combined approach of a semi-structured and integrative approach. A semi-structured approach is used in an arena where multiple disciplines conduct research into a concept and come to different conceptualisations (Snyder, 2019), e.g. used for the notion of AI and safety in chapter 3. An integrative approach is used to illustrate the debate around a topic (Snyder, 2019), e.g. used for the notion of sociotechnical systems and wickedness in chapter 3. The review serves the rigour cycle and as a base for the design cycle in the design science methodology.

As not only a scientific exploration is necessary in order to shape this research, an empirical exploration is conducted as well. The main reasoning is to map the implementations of governmental AI-supported decision-making, leading to the second sub-question:

2. What can be learned about governmental AI-supported decision-making empirically?

This question is answered through a crisp policy and legal analysis in the European Union and The Netherlands, emphasising trustworthy or responsible AI. In answering this question a case study is conducted to enhance the detailed knowledge of the system, as the scientific exploration lacks a detailed answer for the system structure. The goal is to discover the empirical knowledge gaps.

Various opportunities and challenges are defined inter alia through defining the scientific and practical knowledge gaps. Nevertheless, due to the wicked nature of the problem, straightforward solutions are nearly impossible to define. To better understand the system and its inter-dependencies between the contexts, how this influences the outcome, and what interventions can be carried out to make a better system, synthesising research is needed. The system characteristics are defined in the third sub-question:

The second question is answered through an empirical literature review, focusing on official government documents with an integrative approach. The information gathered is primarily used for understanding, conceptualising, and synthesising, which suits the integrative approach (Snyder, 2019). Complementing this methodology is the case study methodology, which identifies the complexity and uncertainties in two cases in the original context of governmental AI decision-making. The first case is elaborated upon extensively, accounting for the influence of law and organisations in 4.4, whereas the latter case in 4.5 is depicted to illustrate the technological context and the lacking implementation of safety measures. Caution is taken into account with generalising the cases directly, warned for by Johansson (2007). Answering this question serves the relevance cycle and as a base for the design cycle in the design science methodology.

3. How can the system and its characteristics be defined?

The system is designed and defined through its characteristics in answering this question. Central are the different contexts of the sociotechnical systems, the actors involved, the definition of the decision-chain and the wicked characteristics in the system. Finally, curiousness is surrounding the safety of the system, which serves as a base for the fourth sub-question:

The third question is answered by synthesising the knowledge from the preceding questions combined with an actors and systems analysis defined by Enserink et al. (2010), including a formal actor chart in which the official relations between actors are conceptualised. Additionally, the system is presented in which the different stages of decision-making are depicted, conceptualised from synthesising

section 3.3, 4.4, and 4.5 considering the theory of sociotechnical systems and the cases.

4. How can citizen's safety be understood over the previously defined system?

The notion of citizen's safety is further explored through ethics and the existing legal base, which further explains the dynamic changes in the system and the boundaries of those changes. Thereafter the final sub-question refers to a semi-empirical approach to understanding system objectives and testing the conclusions of the relationships of citizen's safety. This is explored through the following fifth and last sub-question:

The fourth question builds on the system defined in the preceding question, complemented by a specific literature search to find suitable definitions for a conceptual systems dynamics diagram. System dynamics allows researching the system behaviour under differentiating circumstances to test policy (Keys, 1990). Ideally, this would be helpful for this research as well. However, due to the lack of definitions and quantitative translations from qualitative research, more research is needed to develop a meaningful dynamic model. A different purpose for system dynamics is to understand a part of the real world (Keys, 1990), which is what it is used for in this research.

5. Is the dynamic system behaviour in line with the definition of the system and citizen's safety?

The fifth question tests the preceding two questions. As this field of research is not matured to the extent that it is possible to run a meaningful systems dynamics model, another way of testing and validating the system and its safety is constructed. As the defined citizen's safety is dynamic by nature, an experiment is a set-up consisting of two parts: a serious game and a survey. The serious game is used to assess the decisions made from the perspective of the final decision-maker and suits as it simulates the defined system, measuring the safety levels, which, if elaborated, could serve as a learning objective on citizen's safety as well (Van Der Zee et al., 2012). The game's set-up is extensively addressed in 7.1.

The survey assesses the decisions made and the relation of values toward the system objective. The survey measures the respondents' perception of values over two system objectives, explained in detail in 7.1.

Together, these sub-questions answer the main research question. The following section explains the methodologies used to answer the five questions, where the accompanying research stages are explored, including the iterative nature.

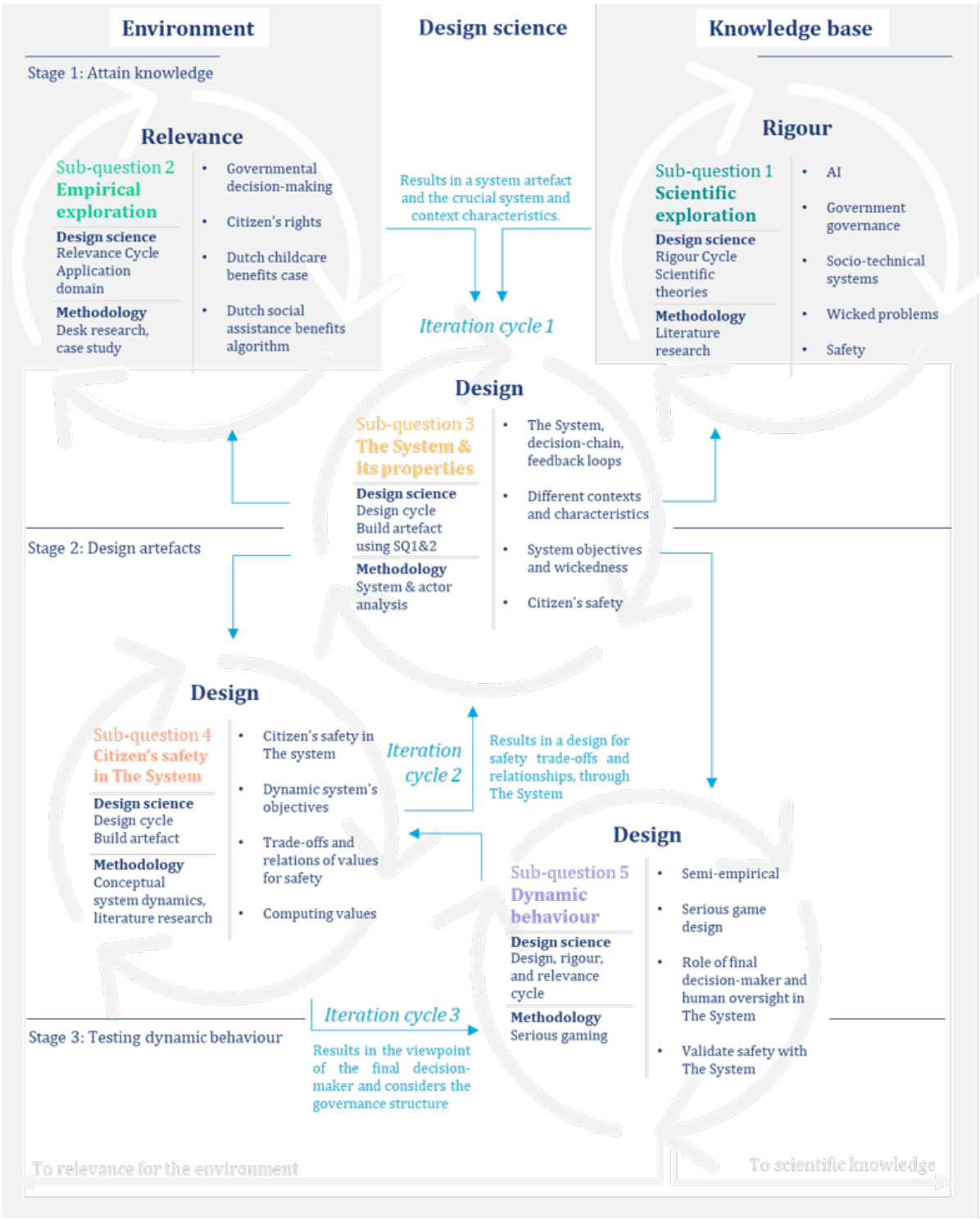
2.3.2. Research stages

This section describes the research stages and methodologies of this research, illustrated in 2.2. The illustration starts with the research stages.

It is paramount to have enough information; as is argued, the knowledge gaps addressed are wide. Therefore, before defining the exact scope, knowledge is gathered. The first sub-question is answered through extensive literature research. The second sub-question is answered predominately through desk research, mainly containing governmental information, and is supported by semi-structured expert interviews. The first question belongs to the rigour cycle, identifying scientific theory. The latter belongs to the relevance cycle, explaining the application domain and arguing the relevance for practice.

After that, a framework is constructed. This research refers to that framework as the system. The system functions as a framework to which the solution can be analysed, which is part of the design cycle through substantiating the relevance and rigour cycle. Execution is done through system analysis, consisting of an analysis of the characteristics and actors. The system can be defined as an artefact, and in reality, it is. In practice, the system is the artefact, or solution, to solve the fraud problem in the case context. The system is discussed and partially validated with experts in unstructured and semi-structured interviews.

Figure 2.2: Research stages



Thenceforth, the so-called solution is constructed to fit the framework. At this stage, the interdependencies between the system and safety and among safety components are reviewed. The challenge is finding a solution that fits all research contexts. System analysis combines literature and desk research to substantiate or prove assumptions. The conceptual model resulting from this stage is validated in the next.

Ultimately, the last sub-question is part of the stage where the conceptual model of the previous stage is examined through a challenging semi-empirical game. The literature lacks empirical research on this topic. Thus this research attempts to add to this knowledge lacuna. A serious game can be seen as design science in itself. However, this research uses the experiment for the rigour cycle, to evaluate and to build the knowledge base. When combining these sub-questions, the main question can be answered.

03 Scientific exploration

"What can be learned from previous research about governmental AI supported decision-making?"

Content

Takeaways

- Undefined functioning for AI systems in government
- Wicked problems contain high complexity, conflict, and uncertainty
- Decision-chain for socio-technical systems
- Lacking safety definition and measurement in system

3

Scientific exploration

This chapter aims to give insight into previous research to reflect upon the scientific knowledge gaps relevant to this research and is conducted through a literature review. The literature review defines relevant notions and places them in the context of this research. It is discovered that definitions are lacking and that the relationship between the notions in this context is unknown. Through the focus points addressed, there is added to the design and rigour cycle, depicted in figure 3.1.

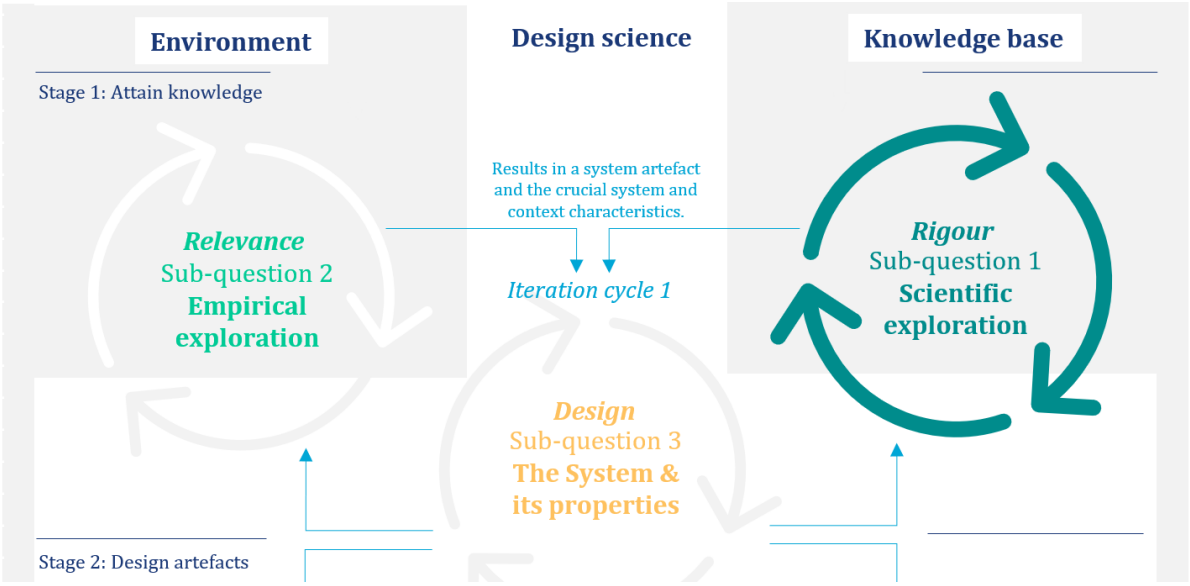


Figure 3.1: Design science cycles relevant for the scientific exploration, inspired by Hevner (2014, p. 88)

Specifically, the sub-question ”What can be learned from previous research about safeguarding governmental AI-supported decision-making?” is answered in the following sections—starting with the literature approach in section3.1, after which relevant notions are discussed. First, the notion of AI is elaborated on in section 3.2, followed by sociotechnical systems in section 3.3, wickedness in section 3.4, and finally safety in section 3.5. This chapter concludes with section 3.7, including the scientific knowledge gaps.

3.1. Literary approach

Most sources are found through Scopus. The challenge in Scopus is finding the right search term. For specific search terms no (relevant) results were found, including searching the title, abstract or keywords, e.g.: *(socio-technical and systems and governmental and ai and decision-making); (”socio-technical system*” and ”Artificial Intelligence” and ”decision-making” and (public or government*)) and*

(limit-to (doctype, "ar")); ("socio-technical" and "System safety" and "decision-making" and (public or government*)) and (limit-to(language , "English")); ("artificial intelligence" and uncertain* and complex* and conflict* and public) and (limit-to (doctype , "ar")) AND (limit-to(language , "English")).*

Some searches only included one source, again searched terms in the title, abstract or key-words: *("artificial intelligence" and "decision-making" and "public values") and (limit-to (doctype , "ar")) and (limit-to (language , "English")); ("artificial intelligence" and government and wicked*) and (limit-to(doctype , "ar")) and (limit-to (language , "English")).*

Or too many to evaluate, again searched terms in title, abstract or key-words: *("artificial intelligence" and "decision-making" and govern*) and (limit-to (doctype , "ar")) and (limit-to (language , "English")); ("artificial intelligence" and government) and (limit-to (doctype , "ar")) and (limit-to (language , "English"))*

Therefore, acquiring the exact research required to answer the sub-question is challenging. To resolve this challenge, Google Scholar is used to search for literature as well. The advantages of this search engine are that it (1) shows the most relevant sources first and (2) has an enormous database. On the contrary, the quality of sources found through Google Scholar might be questionable. Therefore more attention is given to the number of citations, originating journals, and affiliated research institutions.

3.2. Artificial Intelligence

This section generates a mutual understanding about Artificial Intelligence (AI), as the term is ubiquitous. The coverage includes the definition of the notion in section 3.2.1, after which the notion is connected with governmental decision-making in section 3.2.2, after which the last section is presented regarding the different governance approaches defined in the literature in section 3.2.3.

3.2.1. Defining AI

There is no one definition of AI. This research interprets the concept broadly to explore the technology from a multidisciplinary perspective. The typical characteristics include learning from experience and reacting to new inputs, and it often serves for decision-making or prediction purposes (Duan et al., 2019). AI is a developing technology in many fields, from corporate companies to public organisations, and therefore many have researched the topic as well. Makridakis (2017) argue that the future of AI is uncertain, yet the dangers are apparent. However, they argue the difficulty of safeguarding solutions as well. In safeguarding, the difference between private and governmental organisations is that in the case of the latter, the regulator, user and inspector of AI is the same entity: the government. The result of this kind of system is researched in the following chapters.

The lacking definition of AI results in different interpretations in different scientific articles; hence, the opportunities and results are mapped with an explanation of the used definition when it differs from the one used in this research. The definition retained in this study is a system that can learn from experience, react to new inputs, and gives the user information.

3.2.2. Challenges in non-human decision-making

Literature finds different opportunities and challenges for AI. Digitisation and algorithms are additional notions used interchangeably with AI. It is assumed that the challenges and opportunities for digitisation and algorithms apply for AI as well, as the latter only adds to the previous terms with self-learning capabilities. Therefore, this section explains the broad opportunities and challenges for a digitised decision-making government, with and without an AI component.

Agbozo and Asamoah (2019, p. 87) argue in their exploratory study that a data-driven government is "capable of building resilient societies". They define the threats to user privacy & data security breaches, prejudicial biases & labels, and using data-driven decision-making to legitimise powers. These challenges are seen in more articles as well. de Bruijn et al. (2022) identifies seven challenges for explainable AI: lack of expertise; contested explanations, dynamics of data and decisions, interference of algorithms, context-dependency, wicked nature of the problems addressed, and causality is not used for making decisions. These challenges capture why it is not easy for the public sector just to implement any AI.

Levy et al. (2021) identifies three challenges for algorithmic use in the public sector, specifically

after the initiation and during the use of the model, thus at the execution stage. They dive deeper into the technological context. Firstly, they identify that algorithms are not always deployed under the same development conditions. Secondly, they conclude that feedback loops and using output as input data are proven to lead to bias in multiple cases (Ensign et al., 2018). One specific example is the arrest data for drug offences in the United States. The arrest data is used to make decisions about policing, which makes for a 'positive' feedback loop towards black neighbourhoods while resulting in a 'negative' feedback loop in other neighbourhoods, thus creating discriminatory policy (Lum & Isaac, 2016). Thirdly, function creep occurs when the algorithm is not used for the same purpose as it intended. Function creep is defined by Koops (2021, p. 28), as "an imperceptibly transformative and in addition to that contestable change in a data processing system's proper activity". In defining the term, he argues that this terminology makes it possible to hold the right stakeholders responsible and avoid shifting the burden of proof when things go wrong.

Marda (2018) identifies challenges for artificial intelligence through three pillars: data, model, and application. Data is at risk of a lack of access to data when affordable and accurate data is lacking. The research identified such lacking in India, and it is yet unclear how big this challenge is for the Dutch government. Bias in the collection is another challenge because of the risk of overlooking certain communities. Systemic and historical bias, also already illustrated in the article of Levy et al. (2021), can lead to biases against certain groups and create a basis for discrimination.

The risk for models, or decision frameworks, have distinct limitations: feature selection, fairness, transparency and accountability. Data can lead to bias and discrimination, whereas selecting certain features in a model can lead to the same outcome. Feature selection is context-dependent and considered more challenging, as this can occur when safeguards are built. Fairness seems a logical value to handle within modelling but can be very ambiguous. Within models, choices have to be made, for example, between individuals or groups, or accuracy or bias. As fairness can have different definitions in different contexts, it is complicated to make fair modelling decisions. Transparency and accountability are not a given in complex systems, as algorithms can function as a black box. Transparency can be a means to create accountability, and the level of transparency needed can change depending on the context.

Cerquitelli et al. (2017, p. 24) underline the challenges of "violations of privacy, information asymmetry, lack of transparency, discrimination and social exclusion". They argue for three specific requirements to lessen the impact of the aforementioned challenges in the context of social good, centring on humans in its approach instead of technological features. They argue for a positive data-driven disruption through user-centric data ownership and management to ensure privacy, algorithmic transparency and accountability to generate trust, and living labs experimenting with data-driven policies to accept or reject the hypothesis. Altogether, they argue that these requirements will ensure a "data-enabled model of democratic governance running against tyrants and autocrats, and for the people" (Cerquitelli et al., 2017, p. 33).

Choi et al. (2021) argue for more empirical research into the actual implementation of digitised decision-making in the public sector and that, at this moment, developed models are subjective and lack factual data.

The health sector is one public sector that stands out in mitigating risks for citizens using technological innovations. Galetsi et al. (2019) research big data analytics in healthcare. In this research, values such as privacy and security are considered. Furthermore, they argue that the inexplicability of algorithms is a threat to potentially losing medical knowledge. However, they also state that eliminating intuition or personal bias is potentially the most significant advantage. It is noteworthy that the health sector is developing and adapting to new technologies rapidly.

Interestingly, in the aforementioned literature, three challenges are named more than once and include intrusion of privacy, biases, and transparency.

3.2.3. Governance in autonomy

Governing AIs can be done in different ways. In literature, mostly Automated Decision-Making (ADM)s are researched on identifying misbehaviour by including a human. This section examines three types of governance: Human-in-the-Loop (HITL), Human-on-the-Loop (HOTL), Human-in-Command (HIC).

The HITL paradigm is used to address a human handing over necessary information for a machine to use (Amershi et al., 2014; Russakovsky et al., 2015). Rahwan (2018) defines two additional major

functions of a human in a HITL AI system, namely identifying misbehaviour and providing an accountable entity. The system waits for the human to decide to proceed with its operation (Nahavandi, 2017).

Nahavandi (2017) ascribes the additional regulatory functions of Rahwan (2018) to the paradigm of HOTL. Therefore, they argue that this system can continuously run because human only steps in when misbehaviour is detected. These systems require a high level of trust, as the human only monitors and possibly intervenes in the machine's actions (Li et al., 2020; Vierhauser et al., 2021).

HIC is an approach where the human controls the training and adjusting of the model and is a domain expert (Zhu et al., 2018). However, researchers do tend to refer to the concept from the European Union (De Stefano, 2019; Holmberg, 2021; Zhu et al., 2018). Comparing the concept to HITL, the relationship is more of a partnership between human and machine to serve the domain expert, which can be understood as a machine-in-the-loop (Holmberg, 2021).

The concept of human oversight is often build on the assumption that both human and machine have their capabilities, and therefore functions are allocated (Koulu, 2020). When a human is overseeing the performance of the machine, the human is often not able to complete this task, whether due to automation bias, boredom in monitoring, or alert fatigue (Koulu, 2020). The aforementioned arguments are posed in relation to overseeing Automated Decision-Making (ADM). Solutions for the oversight problem can be searched in the design (Almada, 2019; Munir et al., 2013), societal commitment (Rahwan, 2018) changing the technical perspective (Liu et al., 2019), or the way oversight is portrayed in policy (Koulu, 2020). Human oversight in these systems is relatively easy obtained and integrated socially and legally.

However, integrating human decision-makers in the technological context cannot completely, if at all, prevent false outcomes. That oversight or another form of governance is needed is evident (R. A. Smith & Desrochers, 2020). Liu et al. (2019) argue that human rights and the rule of law can be undermined when using machines. This is substantiated by the research of **p.1**Green2022TheAlgorithms, especially for government actors. They argue two major flaws in human oversight of government algorithms because "people are unable to perform the desired oversight functions", and, following from the latter, that "human oversight policies legitimise government uses of faulty and controversial algorithms without addressing the fundamental issues with these tools".

Section conclusion

The definition of AI used in this research is a software-wise technology with self-learning capabilities. It is a broad definition, including predictive modelling in which the variables and its weights are automatically deemed significant and the found patterns applied to a different data set. The definition is chosen to be able to take both the social and technical context of decision-making into account.

The main challenges for AI decision-making are intrusion of privacy, biases, and transparency (Agbozo & Asamoah, 2019; Cerquitelli et al., 2017; Ensign et al., 2018; Galetsi et al., 2019; Marda, 2018). It is substantiated scientifically how the different challenges are affected by the sociotechnical systems, the interventions and their stakeholders. The aforementioned articles focus on specific aspects, however, a sociotechnical systems approach and a holistic perspective are lacking.

Governance for solving the aforementioned challenges is to a large extend based on human oversight, especially in government algorithms, while human oversight is proven to not function as desired.

3.3. Socio-technical systems

This section aims to clarify the definition of a sociotechnical system and how this is viewed among academia. Important to note is the use of a framework in answering the following sub-questions. Therefore, the renowned framework of Rasmussen is presented with several alterations needed to fit the framework to the context of this research.

Sociotechnical systems are defined by Rasmussen (1997) in several layers, research disciplines

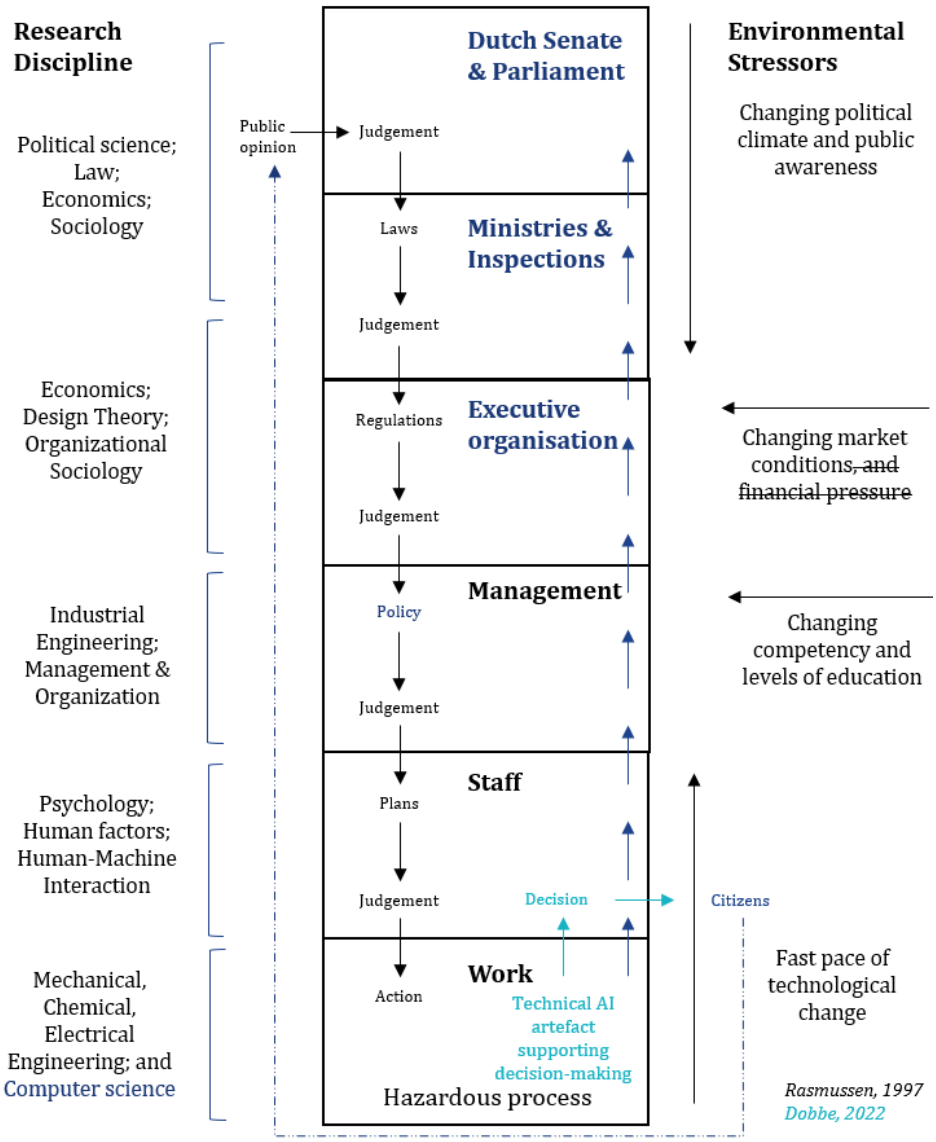
and environmental stressors. The framework is displayed in figure 3.2, and their research is focused on private companies and risk management, which is not entirely in line with this research. In the following enumeration, the changes to the framework are argued.

The definition of sociotechnical AI systems of Dobbe (2022) is added and illustrated in turquoise to add the system components of AI decision-making. Under the last layer, ”Work”, the technical artefact supporting decision-making is added. The work layer influences ”Staff” because it influences their decision-making.

In navy blue, a free interpretation applicable to this research is given. The decision of ”Staff” influences citizens, which is part of the system’s environment. In their turn, citizens can influence politics with their democratic rights. Additionally, at the ”Work” layer, the research discipline computer science is added, as this is the discipline used to create AI. The navy blue also defines the first three layers to adapt to the governmental context of this research. They are specified as the Dutch Senate and Parliament, Ministries and inspections, and executive organisations.

Financial pressure is crossed out due to the self-financing character of the government. Company policy is reduced to policy as it regards governmental policy.

Figure 3.2: Adapted framework for sociotechnical systems towards a public context (Rasmussen, 1997, p. 185)



Eventually, it takes six organisational layers for a problem to pass to the implementation of the solution. The same is true for the feedback upward in the system. Additionally, the people working in the

lower levels are generally regarded differently than those at the top of the organisations. These characteristics might prevent feedback from returning to the dutch senate and parliament. A direct feedback line is noticed between the citizens to which the outcome is influenced and the public opinion choosing who may represent the people in government organisations. This applies by nature to democracies. However, the number of citizens needs to be considered as well. If a solution affects few citizens, the influence on public opinion can be small as well.

Section conclusion

Socio-technical systems are defined by the hierarchal framework of Rasmussen (1997), with the difference being that this research is scoped toward governmental organisations, while the framework of Rasmussen (1997) is scoped toward non-governmental organisations. One relevant insight is that because of the difference of perspective between governmental and non-governmental organisations, laws and regulations set by government are used and weighted differently. For governmental organisations, the laws and regulations behave more dynamically compared to non-governmental organisations, as the laws and regulation deciders are now part of the system, not just its environment.

Eventually, it takes six organisational layers for a problem to pass to the implementation of the solution. The same is true for the feedback upward in the system. Additionally, the people working in the lower levels are generally regarded differently than those at the top of the organisations. These characteristics might prevent feedback from returning to the dutch senate and parliament.

A direct feedback line is noticed between the citizens influenced by the outcome, whom are part of the public opinion, and the public opinion shaping who may represent the people in the (democratic) government organisations e.g. elections, petitions, and is inherent to democracies. Even though this is a direct feedback loop, the question remains whether this is enough vertical integration to assure safe decisions. The reason being twofold. On one hand, the influence of citizens in the public opinion may be questioned, and when the number of negatively affected citizens is low, the shift in the public opinion is meagre.

3.4. Wickedness

This section aims to discuss the scientific consensus on the notion of wickedness to come to a meaningful definition of the term. Vagueness is surrounding wickedness. Therefore, the main components of wickedness are discussed after the portrayed discussion on consensus and its characteristics in section 3.4.1. These are complexity, conflict and uncertainty, of which all are high in a super wicked environment. How these characteristics are defined in this research is read respectively in sections 3.4.2, 3.4.3, and 3.4.4.

3.4.1. Scientific consensus

Ever since the original notion specification from Rittel and Webber (1973), the concept of wickedness has been applied mainly in social sciences, environmental science, business management and accounting. This accounts for more than half of publicised English articles when searching for the query ”wicked” within five words of ”problem” in the title, abstract or keywords, and results in 1651 articles (Scopus, 2022a), of which 969 articles are publicised in the last five years.

The original notion specification is defined as follows (Rittel & Webber, 1973, p. 161-166):

1. ”There is no definitive formulation of a wicked problem (...”;
2. ”Wicked problems have no stopping rule (...”;
3. ”Solutions to wicked problems are not true-or-false, but good or bad (...”;
4. ”There is no immediate and no ultimate test of a solution to a wicked problem (...”;
5. ””Every solution to a wicked problem is a ””one-shot operation”” (...””;

6. ”Wicked problems do not have an enumerable (or an exhaustively describable) set of potential solutions, nor is there a well-described set of permissible operations that may be incorporated into the plan (...”;
7. ”Every wicked problem is essentially unique (...”;
8. ”Every wicked problem can be considered to be a symptom of another problem (...”;
9. ”The existence of a discrepancy representing a wicked problem can be explained in numerous ways. The choice of explanation determines the nature of the problem’s resolution (...”;
10. ”The planner has no right to be wrong”.

Rittel and Webber (1973) started defining the concept of wickedness in social policy by explaining planning problems. The characteristics identified are for this research freely categorised as either problem-focused or solution-focused. Statements one, two, seven, eight, and nine focus on the problem. Statements three, four, five, six, and ten focus on problem-solving. The definition of Rittel and Webber (1973) is firmly examined in the scientific world, especially during the exponential growth from 2006 until 2021 of wicked problems in articles (Scopus, 2022a). In cases critically and cases complementary.

Critics denote the lack of definition and practicality of implementation in the real world (Bannink & Trommel, 2019; Head, 2019; Turnbull & Hoppe, 2019). Some critics connote the ontological difference in social and natural science on which the assumption between the difference between tame and wicked problems is based (Turnbull & Hoppe, 2019).

Consensus does exist on the dimensions of complexity, conflict, and uncertainty of wicked problems (Bannink & Trommel, 2019; Head, 2019; Termeer et al., 2019). Additionally, feedback is a potentially useful tool in this context (Head, 2019; Termeer & Dewulf, 2019).

Termeer and Dewulf (2019) argue for using the concept of small wins to evaluate policy for wicked problems, to prevent the paralysing or overestimating actors when dealing with wicked problems. This approach to wicked problems sheds new light on Rittel and Webber (1973)’ ss paradox of the no-stopping rule, which entails the lacking knowledge of when the problem is resolved when and the requirement of policy actors to judge their strategies. They provide a practical way of evaluating progress while concretely defining goals and making it possible to evaluate the policy continuously. It consists of three steps: identifying small wins, analysing if the proper mechanisms are activated, and finally organising feedback to activate new small wins again. Their research is written from a social sciences perspective and lacks a technical perspective.

Head (2019) is critical of the definition of Rittel and Webber (1973), reasoning that their research lacked knowledge about how to improve policy analysis. They argue that more knowledge about effective policy responses has been gained in the last forty years. The insights can be characterised by conflicts & uncertainties, political complexity, and emerging crises. Concluding, they argue for focusing ”carefully and reflexively on the nature of the policy problems, their evolution, the experience and knowledge of relevant stakeholders and the prospects of effective action in a different situation” (Head, 2019, p. 192). Their research is written from a political science perspective, which explains their focus on political complexity rather than all forms of complexity that emerge in sociotechnical systems.

Turnbull and Hoppe (2019) argue that wickedness should be re-framed towards problemacity, which entails political distance concerning actors because the concept is poorly conceived and, in recent years, has stretched beyond conceptual coherence. They argue for a framework that defines the political distance as delta between (the practices of) each policy worker, where ideas, institutions, and interests influence the problem. This framework allows it to catch both explicit and implicit negotiation and bargaining. However, applying this to a more extensive system of policy workers creates a cluttered overview, and how to practically come to define the difference in problem conception depending on ideas and interests between individual policy workers remains overlooked. Additionally, the framework evolves in the political and cultural realm. Therefore is not entirely compatible with this research. Their research is written from a political and social science perspective, which explains their more combined approach toward problemacity.

Bannink and Trommel (2019) conceptualise the core problem of wicked problems and argue that Rittel and Webber (1973) only observed the problem level emerge. In contrast, wicked problems emerge when at the actor-level, the factual and normative aspects of the issues are intertwined. They argue that

this problem ”is the root cause of the phenomena Rittel and Webber observe: because there is normative conflict and factual complexity, problems tend to be unstable and continuous. That is because the factual and normative dimensions of the problem do not simply coexist; they interact. This interaction of the normative and factual dimensions is, we consider, what enables us to explain the phenomena Rittel and Webber observe” (Bannink & Trommel, 2019, p. 200). They built their research around the concepts of (factual) complexity, which is about the uncertainty of the accurate estimation of social problems, and (normative) conflict, which is about the uncertainty normative evaluation of social problems and solutions. They define wicked problems as ”problems in which there is both a strong difference in the information the regulating and regulated actors have and a strong difference in the values they have” (Bannink & Trommel, 2019, p. 203). Their research is written from a social sciences perspective.

How to deal with and intervene in wicked problems, thus going beyond writing policy and towards implementation, is lacking. Termeer et al. (2019) identify that recent governance approaches have proposed solving wicked problems. In contrast, completely ”solving” a wicked problem is impossible, following the definition of Rittel and Webber (1973). However, they argue that the concept had a negligible impact on policy theories and public policy thinking. They state that, in practice, it either paralyses the policy process or leads to overestimating what can be achieved. Termeer et al. (2019) determine several ways forward for the wicked problem concept. It can be used as a knowledge base to identify failed governance approaches or to define the dimensions of conflict, complexity, and uncertainty more precisely while linking them with contemporary policy science developments. The research of Termeer et al. (2019) is written from a social science perspective. Additionally and importantly, they state that wickedness has recently been used as a buzzword, which is supported by the Scopus analysis previously mentioned, which resulted in an increase of 142% in the last five years (Scopus, 2022a).

3.4.2. Complexity

Complexity is problem-focused, the characteristics one, two, and seven directly add to the complexity because a definite problem formulation and boundaries help to get a grip on the problem. As a wicked problem is essentially unique, every problem must be examined individually. Adding to this are the consequences of system complexity: ”It will often be impossible to disentangle the consequences of specific actions from those of other co-occurring interactions. (...) The outcomes of processes are difficult to predict, amplifying our ignorance and exacerbating the limits imposed by finite resources” (Farrell & Hooker, 2013, p. 686). Conflict and uncertainty contribute to complexity (Bakhshi et al., 2016). Complexity is prone to many definitions in the scientific world. In this research, the definition of systems theory will be taken to examine complex and simple system components, defined by Kinsner W et al. (2010): a large number of interacting elements with many degrees of freedom whose individual behaviour could not be traced back or predicted (p. 277).

3.4.3. Conflict

Conflict is introduced through a multi-actor system, where actors lack common ground, implicitly or explicitly, in solving the problem. Conflict can be expressed in a normative way through differences between values and norms among actors. These actors are intertwined with the problem and resolution. (Farrell & Hooker, 2013). Characteristics three, four, and five link wickedness and conflict. As there are no true or false solutions for wicked challenges and no test for solutions, actors can differ in opinion about policy or implementation. Additionally, the chance for compromises among actors decreases when solutions are sought, as solutions influence the problem. Therefore, the concept of conflict becomes relevant when solutions are sought.

3.4.4. Uncertainty

Uncertainty can be explained in model-based decision-making through seven levels of uncertainty. W. E. Walker et al. (2013) define the levels from 0, complete certainty, to 6, total ignorance. In complete certainty, the chance of some event is 100% and for total ignorance 0%. The other levels are between levels 0 and 6 and can be explained through the future world of a policy problem: ”Level 1: A clear enough future (with sensitivity); Level 2: Alternate futures (with probabilities); Level 3: Alternate futures (with ranking); Level 4: A multiplicity of plausible futures (unranked); Level 5: Unknown future” (W. E. Walker et al., 2013, p. 4). Levels four and five are categorised as deep uncertainty. Uncertainty is seen

through the lacking immediate and ultimate solution for a wicked problem. Furthermore, the uniqueness of a problem contributes to its uncertainty. Therefore, uncertainty can occur both at the problem and the solution.

Section conclusion

Complexity, conflict, and uncertainty are central concepts that recur in literature and contribute to a better understanding of wickedness, as consensus about a precise definition and practical use is lacking. These concepts need not be described as binary in an ”is” or ”is not” but rather be seen at a continuous scale, where the system has a certain level of wickedness, yet system components might have differences. These concepts are used throughout this research.

The reason for using wicked problems as a concept differs among academics (Lönngren & van Poeck, 2021). The rhetorical function of this research is to both challenge existing approaches and support alternative approaches. The first through challenging ”the dominance of a specific group in addressing [the] problem” (Lönngren & van Poeck, 2021, p. 490) and the latter through the ”usefulness and value of a specific scientific discipline, (...) [and a] call for action within a specific social community” (Lönngren & van Poeck, 2021, p. 491).

Knowledge lacks on the behaviour of wicked problems when confronted with an operational challenge, as the aforementioned literature shows a policy perspective on the notion. The difference between the two perspectives is that in policy the problem is prone to many other decisions after policy is determined. For an operational wicked problem, the last decision is made in the operational stage before its impact flows out of the system toward its environment.

3.5. Citizen’s safety

This section elaborates on safety. Firstly, the safety perspective is elaborated in section 3.5.1. Thereafter safety and its continuity are explained in 3.5.2. After that, the concept of safety culture is introduced through shared values in **??**. The final part of this section refers to the concepts of resilience and robustness and their use for safety in 3.5.3.

3.5.1. Safety in a system perspective

System safety engineering is defined by N. G. Leveson (2011, p.468) as: ”The system engineering processes used to prevent accidents by identifying and eliminating or controlling hazards. Note that hazards are not the same as failures; dealing with failures is usually the province of reliability engineering”, which creates curiousness into the definition of accidents and hazards. Accidents: ”An undesired and unplanned event that results in a loss”, and hazards: ”A system state or set of conditions that, together with a particular set of worst-case environment conditions, will lead to an accident (loss)” (N. G. Leveson, 2011, p.4867).

The concept of system safety is born in industrial and computer science yet is becoming more relevant to a broad public through the intertwining between technology and society (N. G. Leveson, 2011). The boundaries of the system depend on the context. New assumptions are set by N. G. Leveson (2011), to fit the frequent complex systems:

1. ’High reliability is neither necessary nor sufficient for safety’ [p.14]
2. ’Accidents are complex processes involving the entire sociotechnical system. Traditional event-chain models cannot describe this process adequately’ [p.31]
3. ’Risk and safety may be best understood and communicated in ways other than probabilistic risk analysis’ [p.36]
4. ’Operator behaviour is a product of the environment in which it occurs. To reduce operator ‘error’ we must change the environment in which the operator works’ [p.47]
5. ’Highly reliable software is not necessarily safe. Increasing software reliability or reducing implementation errors will have little impact on safety’ [p.50]

6. 'Systems will tend to migrate toward states of higher risk. Such migration is predictable and can be prevented by appropriate system design or detected during operations using leading indicators of increasing risk' [p.52]

7. 'Blame is the enemy of safety. Focus should be on understanding how the system behaviour as a whole contributed to the loss and not on who or what to blame for it'[p. 56]

These new assumptions set light to the meaning of a safe system in sociotechnical systems. The work of N. G. Leveson (2011) builds on the work of Rasmussen (1997) on sociotechnical systems (N. G. Leveson, 2017). Generally, Leveson applies their ideas to roughly three subjects: workplace safety, aviation, and medicine (Scopus, 2022b). All are sociotechnical systems: people coming together with technology. Applying the system safety concept to the topic central in this research, medicine is most similar as the safety of a non-decision-maker human, the patient, is crucial. Therefore, the safety control structure is often elaborate, as illustrated in N. Leveson et al. (2020).

N. G. Leveson (2011) present the foundations of systems theory emergence & hierarchy and communication & control. Emergence & hierarchy refer to the emergent properties of the system which do not exist on lower levels of organisation; the levels referring to the hierarchy of the system, generally referred to in complex systems. How communication and control is shaped is relative to the level of hierarchy. Constraining the activity on one level defines as control, and when the system contains input of and output to their environment, communication is required. For controlling a process for safety, four conditions are required (N. G. Leveson, 2011, p.65):

1. "Goal condition: the controller must have a goal or goals

2. Action condition: the controller must be able to affect the state of the system

3. Model condition: the controller must be or contain a model of the system

4. Observability condition: the controller must be able to ascertain the state of the system"

The four conditions for safety processes lead to designing for safety. The conditions lead to the system changing from an unsafe state to a safe state (the goal) through the actions of a controller, who is able to oversee the system's state. Before discussing the four conditions, safety must be defined to assign accompanying goals. To discuss on safety processes, the system, its boundaries, and the relation among system components must be defined, besides safety.

In straightforward systems, i.e. a pilot steering a passenger aircraft, one safety condition may be defined that the pilot can lands safely after take-off. Safety processes may be installed, e.g. the cockpit is locked, so passengers cannot open it in the event of a hostile take-over. In this socio-technical system, safety is defined as no accidents, and accidents are defined as the occurrence of loss (N. G. Leveson, 2011). The accidents are often visible to the stakeholders and/or bystanders, e.g. the crashing of the plane is mostly directly noted. Safety issues come from coupling effects of components and are non-linear because of feedback and coordination, and therefore the system knows dynamic behaviour as well (N. G. Leveson, 2011). Threats to safety can come from a lack of vertical integration and system defences erode over time. The characteristics of this system safety approach are the complexities and uncertainties under which accidents may occur, including the aforementioned.

3.5.2. Continuous safety

On the other hand, reliability refers to the not failing of the system (components) under given circumstances, yet does not guarantee a safe outcome, meaning that safety has to be looked at from a system's view. N. G. Leveson (2011) identifies another assumption regarding safety in a complex sociotechnical context, which refers to considering the system component's role and interaction to understand and place an individual component in the system. Furthermore, they state that "bottom-up decentralised decision making can lead - and has led - to major accidents in complex sociotechnical system" (p. 14) because it can lead to accidents when the decisions and behaviours of the organisation interact dysfunctional.

System safety methodologies are broad and traditionally focused on a root cause through a sequence of events. However, the paradigm is shifting toward a more holistic approach because technology has become more complex and rooted in a social context. Since the 90s, methods like Swiss

Cheese, AcciMap, STAMP, CREAM, and FRAM have been used in the context of complex sociotechnical systems (Waterson et al., 2015). According to Waterson et al. (2015), most methods represent the environment and context of the system in a weakly manner. Additionally, the political, legislative or regulatory factors are left out by most methods. Therefore, this research looks into safety in another manner.

System safety can be built by design as one of the initial steps in creating the system. However, public sociotechnical systems are often already designed, and sub-solutions are provided for the more minor challenges, therefore dissociating them and disregarding the line of arguing of N. G. Leveson (2011). While safety by design is essential, this research focuses on continuous system safety, as the previous section emphasises feedback and continuous evaluation in a wicked system.

3.5.3. Resilience & Robustness

Resilience and robustness are popular concepts in many fields, including policy design, and are used as solutions to govern complexity (Capano & Woo, 2017). Resilience refers to the ability of a system to return to its equilibrium after a shock. For policy design, Capano and Woo (2017) argue that this might mean that the core of the shock was not sufficiently resolved, and thus adaptation might be more desirable. On the contrary, resilience also knows a dynamic interpretation, allowing flexibility and adaptivity (Tanner et al., 2017). However, interpreting resilience with a dynamic property might better be referred to as robustness (A. Smith & Stirling, 2010). Robustness refers to the ”ability to withstand or survive external shocks, to be stable despite uncertainty” (Bankes, 2010, p. 2). Capano and Woo (2017) argue that robustness is valuable for policy design, as the nature of the concept allows for an adaptive system, not having to return to the same equilibrium before as after a shock. Both concepts can be used to operate under uncertainty and complexity (Bankes, 2010), which bridges the gap between wickedness and system safety.

Resilience and robustness are considered in the system safety of a sociotechnical system. This research defines the concepts as they can be applied to many fields. Rather than naming them as the goal, they contribute to the ability to write about the system and its safety. System safety is more than a resilient and robust system. It might be contradictory in certain parts of the system.

Section conclusion

System safety is an established notion among academics, which primarily entails a technical perspective. Socio-technical systems are considered, i.e. aviation. Applying the terminology to a societal wicked problem requires the definition of safety, which can be a challenge. In this, macro perspective in which the system is considered politically may collide with the level of specificity required to analyse the system according to system safety in sociotechnical systems. For the political and organisational perspective, including the actor playing field and the shaping of the rules of the game, a macro or meta perspective fits, while for the technological and more operational organisational perspective the notion of system safety fits, which is inherently from a micro or, in cases, a meta perspective. This is the crux of the way of analysing in this research: combining both a specific algorithm’s system with a complex actor playing field in which rights and functions know implicit characteristics.

3.6. Cohesion among notions

This chapter elaborates on the paramount notions in this research. Combining the knowledge portrayed, it is possible to examine the knowledge gaps in-depth. The found knowledge gaps capture a broad opening to which the general conclusion can be stated as that the concepts are not yet interpreted in context to each other.

Safety in engineering defines as minimising risk and steering clear of hazards. In engineering, the risks and hazards are often straightforward. For example, preventing a fire can be done by removing oxygen, fuel or a detonator. Suppose the fire occurs anyhow through a compromise in the aforementioned. In that case, the impact of the fire can be decreased when safety measures are implemented, for example, having a fire extinguisher close by the high-risk locations. When applying safety to wicked problems, the solutions might not be as straightforward as having a fire extinguisher. The commonality is that also,

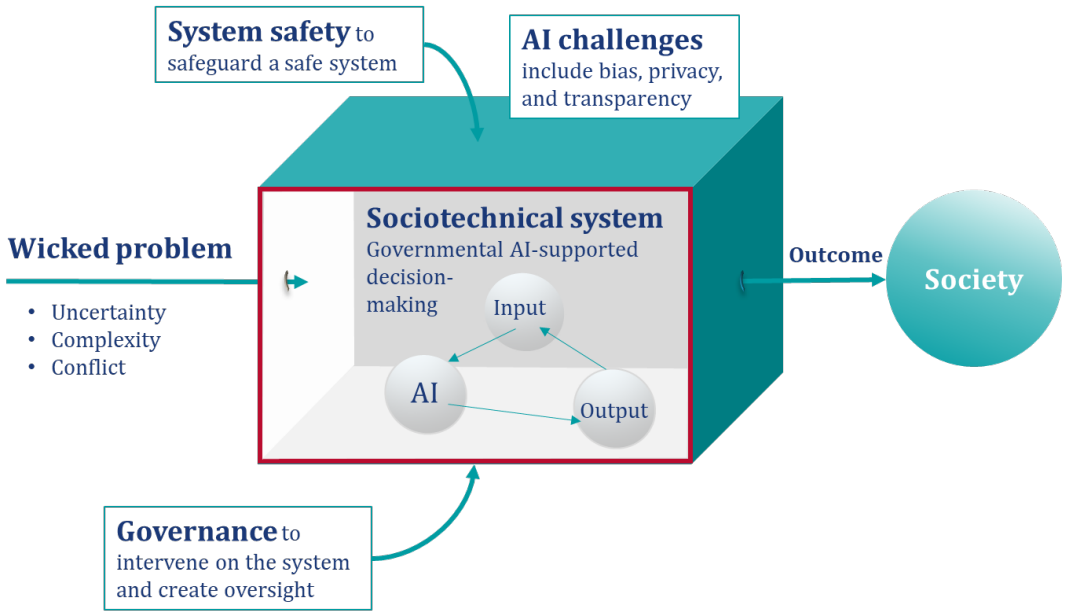
in this situation, the will is to prevent accidents from occurring. Differences lie in the solution that can be applied, which can be explained through the wicked environment and the problem solved in the system.

Wickedness refers to the complexity, uncertainty and conflict in governmental decision-making processes combined with the characteristics defined by Rittel and Webber (1973). For this research's wickedness, crucial factors are that solutions to wicked problems are not true-or-false, but good or bad, that there is no immediate and no ultimate test of a solution to a wicked problem, and that the choice of explanation determines the nature of the problem's resolution, and, lastly, that the planner has no right to be wrong. The planner, in this context, ultimately means the final decision-maker. An ultimate definition or solution is not available. Therefore, wickedness portrays the complex and uncertain environment in which conflict may occur. Wickedness makes it possible to discuss the aforementioned characteristics and understand why a solution is neither easily acquired nor implemented. Wickedness also creates the need to continuously review what is happening in the system, whether the solutions are good or bad. Thus a continuous and not a static feedback process is desired..

Governance is interpreted broadly, containing the control and/or oversight structures in the system. Ideally, the system's outcome is completely correct, to which governance structures can steer. As it is not yet known when this requirement is fulfilled in the context of a wicked problem, how the political, organisational, and technical components ought to be designed is unclear; hence, is not substantiated how to intervene in the system. Currently, as apparent in the following chapter, chapter 4, the human-on-the-loop governance structure controls the model's output before it addresses society. Unfortunately, it is unclear how this is precisely done in practice and what intervention space is given to these final decision-makers.

Combining these notions presents the angle of this research more clearly and is summarised in 3.3. The problem at hand is wicked, and the system to find a solution for the wicked problem consists of the system illustrated as the cube. The outcome affects citizens in society, which creates a high impact when the wrong decisions are made. Designing governance for the system makes it possible to intervene for citizen's safety. Especially feedback loops can be enabled by suitable governance structures.

Figure 3.3: Cohesion among notions



There are many unknowns, and uncertainties regarding governmental AI-supported decision-making, especially when this holistic perspective is taken into account. This research further examines the system, its components, and internal interdependencies, intending to guide towards a safe, responsible,

and trustworthy system where decisions can be made more substantively throughout politics, organisations and technologies.

3.7. Conclusion & scientific gaps

This chapter gives insight into previous research to reflect upon the scientific knowledge gaps relevant to this research and is conducted through a literature review. Through the literature review relevant notions are defined and placed into context. This section concludes the literature review and presents the scientific knowledge gaps. It is discovered that the relevant notions for this research have not jointly been researched in a direct manner. To come to this insight the sub-question ”What can be learned from previous research about safeguarding governmental AI-supported decision-making” is central. Key concepts are defined to answer this question, including sociotechnical system, wickedness, safety and Artificial Intelligence (AI).

Many scholars have defined the concept and researched specific parts of Artificial Intelligence (AI). For instance, black-box characteristics, transparency, privacy, and bias are all researched, regularly from a different perspective. In the public sector there are both academics who use a technical approach, and those who use a more social approach to AI (Vydra & Klievink, 2019). The challenges of transparency, privacy, and bias are known, however, how they relate to each other depending on the part of the system they relate to is not. Therefore, the aim of this research is to gather these challenges and provide insight into their relation to the system.

Knowing the challenges of AI steers in a direction for the consideration of safety. Safety is thoroughly described for sociotechnical systems by the system safety doctrine, i.e. by the work of N. G. Leveson (2011). Crucial is that it is known what is safe and what is not. For some systems safety can be unambiguous, i.e. when a fire in a factory kills its workers, the working conditions are unsafe because an accident (loss) has occurred and can be identified immediately. How the concept of safety and system safety can be applied to wicked problems with a political component is out of scope in the current system safety doctrine. N. G. Leveson (2011) builds on the work of Rasmussen (1997) regarding sociotechnical systems.

Sociotechnical systems are defined by having social and technical components. Rasmussen (1997) defines a sociotechnical system hierarchically, stepping from a macro to micro perspective. The definition of Rasmussen (1997) can be used for this research with two alterations. First, the sociotechnical system is defined for an organisation or company. However, the scope of this research are the governmental decision-makers. Therefore, the political context is reflected in the system rather than an external factor. This perspective is explained in the following chapter, chapter 4, and the consequences become evident from chapter 5 onward. Secondly, the final decision is not made at the last and technical stage, in which the fast pace of technological change is evident, but goes one hierarchical step up. The reason being that the technology supports the decision-making process, in this research through AI. Lastly, the only feedback loop existing is from citizens impacted by the decision toward the public opinion. This feedback loop is large and, hence, vertical integration is lacking. Whether other feedback loops are present is to be researched.

Wickedness is prone to a discussion in the academic world, where scientists argue over the goal of use for the notion ever since Rittel and Webber (1973) defined the notion according to their characteristics. Wickedness is generally defined from a policy perspective, lacking operational function. Complexity, conflict, and uncertainty are central concepts that recur in literature and contribute to a better understanding of wickedness, as consensus about a precise definition and practical use is lacking. These concepts need not be described as binary in an ”is” or ”is not” but rather be seen at a continuous scale, where the system has a certain level of wickedness. Wickedness makes for a challenging arena to intervene in, potentially leading to stakeholder conflict. This research uses the notion of wickedness firstly for defining the problem and showing the complexities, potential conflicts, and uncertainties and on this sum academics do agree. Additionally, wickedness used to evaluate the trade-offs occurring at the operational stage.

Combining a wicked problem with safeguarding the safety of the socio-technical system creates a situation which is not addressed in science, yet the reality for many governmental practitioners. As research is lacking, an attempt to fill the gaps is conducted in this research. To do this, two more notions are crucial to debate the topic. First, the meaning of safety ought to be defined, which fits the social

and technical components. Second, a manner in which all contexts of a sociotechnical system can be evaluated is desired, including social and technical aspects.

Research into the preceding literature on safeguarding governmental AI-supported decision-making resulted in the consideration of the notions AI, wickedness, sociotechnical systems, and safety. How the notions are interconnected is clarified in the aforementioned section, cohesion among notions. The notions are related through defining the problem with the help of wickedness, for which a sociotechnical system is made as a solution. The impact of the decisions made in that system, including the technological AI, impact the citizens in society. Ideally, the system is safe, and measures to safeguard its safety is implemented. In this way, the challenges regarding AI-supported decision-making can be deflected.

04 Empirical exploration

"What can be learned empirically about governmental AI supported decision-making?"

Content

Takeaways

- The Netherlands is frontrunner in decision-making AI
- Equality, privacy, and transparency are central in the AI policy of the European Union
- Wickedness and need for safety is empirically proved through the childcare benefits case
- Measures for safekeeping are explored through the social assistance benefits case

4

Empirical exploration

This chapter aims to give insight empirically to reflect upon the empirical knowledge gaps relevant to this research and is conducted through desk research. The desk research ought to define the scope and applied cases. It has been discovered that the Netherlands is the global leader in decision-making AI, and policymakers want trustworthy AI. However, the notion is not explicitly defined, and even though the previous chapter 3 shows the scientifically unsubstantiated base, in the real world, such algorithms are operational. Through the focus points addressed, there is added to the relevance and design cycle, depicted in figure 4.1.

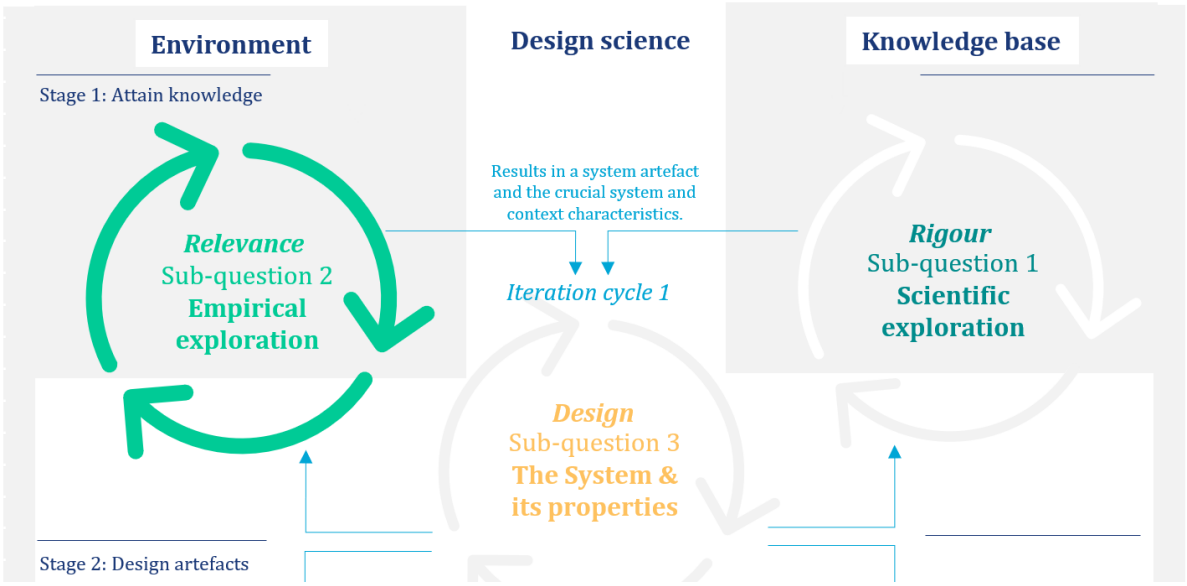


Figure 4.1: Design science cycles relevant for the empirical exploration, inspired by Hevner (2014, p. 88)

Specifically, the sub-question ”What can be learned empirically about safeguarding governmental AI-supported decision-making?” is answered in the following sections. Starting with global AI developments in section 4.1, followed by the notion of trustworthiness in section 4.2. After that, the case selection is elaborated upon in section4.3, continued by both the elaboration of the childcare benefits in section 4.4 and the social assistance benefits case in section 4.5. Finally, the conclusion and empirical knowledge gap are depicted in section 4.7.

4.1. Global AI developments

Other democratic countries similar to the Netherlands are implementing AI as well. For example, Canada, Japan, and the United States developed strategies and created principles for a human-

centred, responsible or trustworthy applications (Government of Canada, 2021; Ministry of Internal Affairs and Communication, 2019; United States government, n.d.-b). Germany has not yet come to a regulatory framework, but did develop a strategy and already knew many applications used by government (German Federal Government, 2020; Lernende Systeme, n.d.). Comparing the tools and frameworks clarifies transparency, privacy, equality, robustness, and fairness as central measurement nodes; however, it remains unclear how to measure precisely, other than factoring in or mitigating the risks (Government of Canada, 2022; United States government, n.d.-a). What is lacking in these countries is detailed public information concerning the algorithms and the socio-technical system.

Furthermore, international initiatives exist.Digital Nations, G7, G20 all concern a global digital collaboration, of which AI is part (Government of Canada, 2021; United States government, n.d.-c), and the European Union (EU) Union is closely engaged as well. They published ethics guidelines, introduced a framework, and defined four principles: respect for human autonomy, prevention of harm, fairness, and explicability (High-Level Expert Group on AI, 2019). Research from the Rathenau Institute shows that the Netherlands is the leader in decision-making applications of AI, compared to front-runners like the United Kingdom and the United States (Koens & Vennekes, 2021).

This research specifically chooses the Netherlands as its geographical boundary, as the Netherlands is a democratic country in which the federal government can be held accountable, a leader in decision-making AI and sufficient documentation is publicly available.

4.2. Trustworthy AI

Responsible, trustworthy, and safe AI is dominating the policies, as seen in the previous section. This section defines the ethics and legality surrounding AI in the Netherlands and European Union (EU).

The European Union defines different requirements that need to be to be able to fulfil trustworthiness (High-Level Expert Group on AI, 2019, p.14):

- Human agency and oversight
- Technical robustness and safety
- Privacy and data governance
- Transparency
- Diversity, non-discrimination and fairness
- Societal and environmental well-being
- Accountability

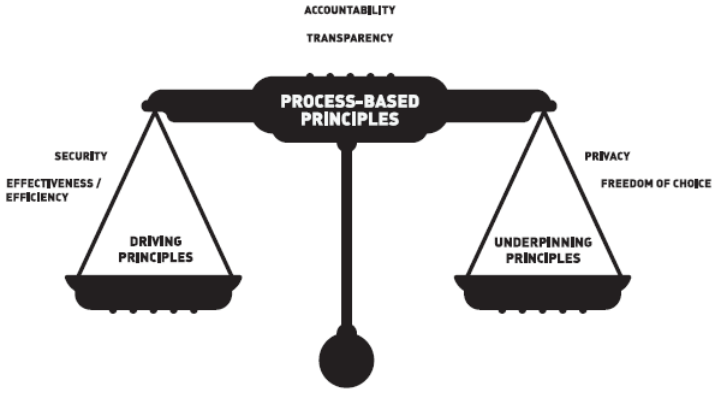
Additionally, they show critical concerns regarding enabling citizen scoring and argue it violates fundamental rights and can only be conducted if a clear justification and measures are fair and proportionate. What this explicitly entails remains unclear.
Dutch Scientific Council (WRR) conducted research about digital government and public values (Prins et al., 2011, p.66). They identified three categories of values: driving principles, underpinning principles, and process-based principles, illustrated in figure 4.2. Driving principles represent values like effectiveness and efficiency, which is why it is useful to implement technology. Underpinning principles represent values like privacy and fairness and might come under pressure with implementing technology. Process-based principles represent values like accountability and transparency. These principles are used to gain insight into the system. Additionally, the values can be balanced, hence the illustration of the scale.

4.2.1. Legal basis for ethics

For the protection of citizens against government, values are embedded in law. Equality, privacy, and transparency are part of the requirements for trustworthy AI according to the EU. Equality is the term this research uses for non-discrimination and fairness as defined by the EU.

Equality is embedded in the constitution, as well as in the Equal Treatment Act, the Equal Treatment of Disabled and Chronically Ill People Act, the Equal Treatment in Employment (Age Discrimination)

Figure 4.2: Tripartite division of principles (Prins et al., 2011, p.66)



Act, and the Equal Treatment (Men and Women) Act (Government of the Netherlands, n.d., 2018). The laws prescribe equal treatment and prohibit discrimination based on race, sex, sexual orientation, age, nationality, etcetera. The opposite of equality is often called bias. Different meanings relate to the concept from a technical, socio-technical, and societal perspective (Paola, 2021), where technical bias refers to the deviation of data through the model, socio-technical bias refers to the deviation due to inequalities, and societal bias refers to the inequalities in society. The latter is depicted in law.

Transparency is anchored in law by the Open Government Act (Ministry of Internal Affairs & Ministry of Justice and Safety, 2022). This law regulates the openness of the government, obligating them to be transparent actively. Part of transparency is the algorithm registers publicised by some public organisations and are only used by the four large municipalities: Amsterdam, Utrecht, The Hague and Rotterdam. However, they do not cover all the (publicly known) information of said algorithms and are not all as mature nor complete (Algoritmeregister, n.d.). Besides transparency of algorithms, the Dutch government also works on open data initiatives, of which more than 15.000 are publicly available (Dutch Government, n.d.).

Privacy has been embedded in Dutch law since 2018 by the General Data Protection Regulation (GDPR). The law gives the people several rights: to be informed, to access, to rectification, to erasure, to restrict processing, to data portability, to object, and rights concerning profiling and automated decision-making (General Data Protection Regulation (GDPR), 2016; Wolford, n.d.). The latter is of explicit importance for this research. Under specific circumstances, profiling and automated decision-making are allowed, which includes a legal basis and a privacy impact assessment, and thus takes care of the need for human oversight (General Data Protection Regulation (GDPR), 2016, Art. 22), and showcases privacy by design principle.

4.2.2. Human oversight

Human oversight ”may be achieved through governance mechanisms, such as a human-in-the- loop (HITL), human-on-the-loop (HOTL), or human-in-command (HIC) approach” (High-Level Expert Group on AI, 2019, p. 16). In 3.2.3, the scientific definitions were already illustrated, including the conclusion that the EU is often connected with the notion of HIC. In the ethics guide of the EU, the paradigms are defined slightly different (High-Level Expert Group on AI, 2019, p. 16).

Firstly, HITL refers to ”human intervention in every decision cycle of the system” High-Level Expert Group on AI (2019, p. 16), which does not empathise with the need for the human to hand over crucial information to the machine and creates the idea of intervention on the system instead of the human being crucial in the system.

Secondly, HOTL refers to ”the capability for human intervention during the design cycle of the system and monitoring the system’s operation” (p. 16), where the use of intervention is in line with the scientific definition. However, crucial and lacking is the notion that for the human to intervene, it first has to recognise the misbehaviour of the machine, which adds complexity to the approach.

Lastly, HIC refers to ”the capability to oversee the overall activity of the AI system (including its broader economic, societal, legal and ethical impact) and the ability to decide when and how to use the system in any particular situation” (p. 16). As defined in science, HIC refers to the command and

control of training data and the adjustment of the machine by a human domain expert. Even though human intelligence can be intelligent, it is magnified if one states that humans can oversee **all** overall activity of a system. A domain expert may not always oversee the multidisciplinary environment and indirect correlations.

4.2.3. Remaining requirements

The remaining requirements, technical robustness & safety, societal & environmental well-being, and accountability, are equally important.

Resilience and robustness are important in the security of the system. Additionally, according to the EU, the system's accuracy is part of technical robustness & safety. Specifically, they state that "a high-level accuracy is especially crucial in situations where the AI system directly affects human lives (High-Level Expert Group on AI, 2019, p. 17). It is unclear if this section also refers to automated or augmented decision-making. Relative accuracy in this research is defined as effectiveness.

In societal & environmental well-being, it is stated that the effects on citizens should be carefully considered, monitored, and assessed from a societal perspective and is in line with the principles of fairness and the prevention of harm.

Lastly, accountability is also required for trustworthy systems and relates to the principle of fairness. They state that due protection must be available if an entity reports legitimate concerns. Impact assessments can help reduce the negative impact during the design and operating stages. Additionally, trade-offs should be explicit, and the decision-maker must be accountable for them.

4.2.4. Methods

Both technical and non-technical methods are recognised. In this part, the methods identified by High-Level Expert Group on AI (2019) are elaborated on.

Technical methods

- Architectures for Trustworthy AI
- Ethics and the rule of law by design
- Explanation methods
- Testing and validating
- Quality of Service Indicators

Non-technical methods

- Regulation
- Codes of conduct
- Standardisation
- Certification
- Accountability via governance frameworks
- Education and awareness to foster an ethical mindset
- Stakeholder participation and social dialogue
- Diverse and inclusive design teams

A systemic, continuous and multi-value approach is not listed above. A combination of the methods might create such an overview, yet it is not the main focus of the methods. Technical and non-technical methods are mentioned, yet socio-technical methods are absent, and the political perspective is not included. Additionally, according to them, it lacks a method for explicating trade-offs, which is at the core of trustworthy AI.

4.3. Case selection

With the aim of the research in mind, it becomes evident that it is necessary to go into more detail and specification than can be defined generically. Therefore, this research specified a case that needs to fulfil defined requirements. First, it should fit the research goals and therefore be a system to solve operational challenges with AI-model(s) in a governmental context, impacting citizens. To make a useful analysis of a said the system, one needs to be able to find the right details of the system. Furthermore, this research will be publicly available, which requires open(ing up) information. These requirements

are fulfilled by one case that has been under public scrutiny: the childcare benefits affair, which resulted in a lot of harm to citizens trying to make ends meet. Therefore, the childcare benefit case is researched in detail in this section.

In some parts, a comparison to other cases suits to aim for more general conclusions. The difficulty is that no other case has as much publicly available information as the childcare benefit case. Argumentation entails the importance of the secretive nature of the detection of criminals. Another case that uses a similar system is the social benefits case used in the municipality of Rotterdam.

One issue for which AI is used in many countries is the COVID-19 virus to adapt better policies and is purposefully not opted for, as the models and data for the virus are still adapted and gathered. It is not evaluated to have harmed citizens extensively.

4.4. Childcare benefits case

The childcare benefits affair affected 63.120 children negatively by demanding repayments (CBS, 2022b), between 2013 and 2020 when the risk model for profiling was in use (Autoriteit Persoonsgegevens, 2018; Belastingdienst, n.d.), and even resulted in 1.675 out-of-home-placements (CBS, 2022b). This case is debated nationally, and internationally (Henley, 2021b; Holligan, 2021; Roobeek et al., 2021), and Jesse Frederik even wrote a book about it (Frederik, 2021b).

In this case, the government's aim is elaborated on in 4.4.1. Many things have changed in society, politics, organisations, and technology. The timeline of the case is reconstructed in 4.4.2 to capture the changes over time according to law and organisation. Lastly, the process parents had to go through to receive the benefit is illustrated in 4.4.3.

4.4.1. Aim of government

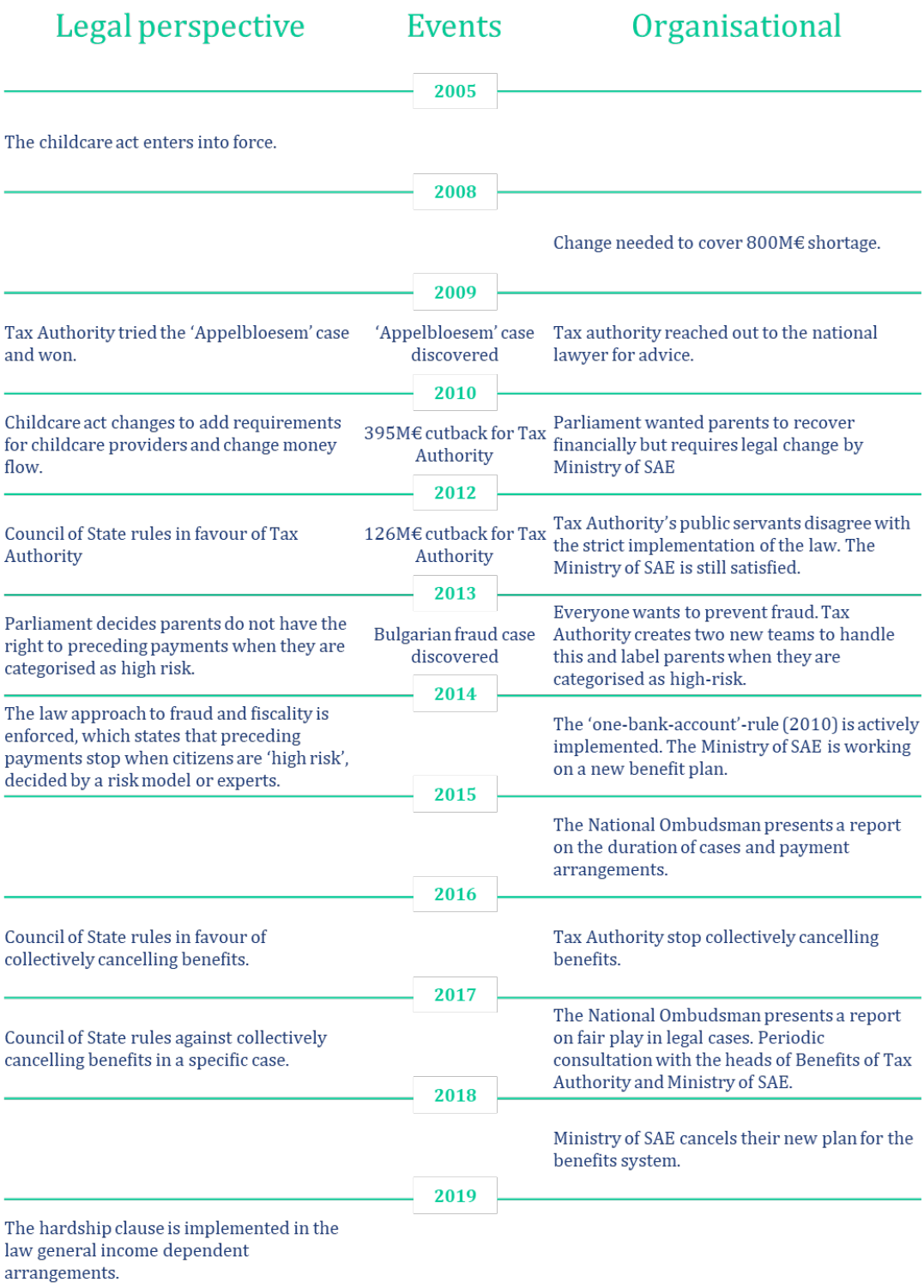
Although the government is paved with many responsibilities, the research realm specifies that the goal of government is threefold: (1) they want to take care of citizens who need it, (2) prevent citizens are taking advantage of the law, and (3) obey the law themselves (Overheid.nl, n.d.; Rijksoverheid, n.d.-a). In the case of fraud detection, the government chooses to categorise citizens through predictive modelling, training a model with (personal) data, requiring high-quality models and data, and aiming for a rightful outcome. The outcome is theoretically rightful if there are no false positives and no false negatives. Because of the consequences the outcome can have on one's personal life (Autoriteit Persoonsgegevens, 2018), false positives cannot be accounted for in a fair government. However, false negatives, meaning letting people take advantage of tax-payer money, are unwanted in a fair government as well. In general, the government's goal is to take care of citizens while obeying the law and preventing others from taking advantage of it. In the described benefits case, the goal is to provide benefits fairly without causing harm to citizens.

4.4.2. Timeline elucidation

In 2013, the Dutch House of Representatives voted unanimously in favour of a new law, "Wet aanpak fraude toeslagen en fiscaliteit" (Weekers F.H.H. & Opstelten I.W., 2013), the law approach to fraud in benefits and taxation. This law allows risk modelling to single out citizens, check benefits requestors beforehand, deny benefits when a requester is unknown, and take more time to check benefits. In 2014 the law was entered into force. The only entity critical of this law and its rapid development was the **CoS!** (**CoS!**) (Council of State, 2013), the independent advisor of the Dutch government. This law is followed up with a 25 million euros investment, and according to the fraud team at the tax authorities, every euro spent would return 3,30€. It would result in a profit of 57,5 million euros. If not enough benefits were shortened, the tax authorities would have to cut from their organisation (Amnesty International, 2021). For the tax authorities, a new legal approach to cut benefits emerged, and a financial incentive to execute a cut was implemented, emerging from politics. Besides financial incentives to detect fraud, the tax office had to implement more laws and changes to laws. However, the problem's origins started in 2005, when the first version of the law was enforced, and the first problems started in 2010 (2Doc, 2021), years before the Bulgarian fraud case was discovered in 2013 (Steinglass M, 2013). The law and organisational perspective are captured in the timeline in figure 4.3.

2005 The childcare benefit regulation is set up to reduce costs and increase quality. The law states that the parents must agree with the childcare provider and pay the bill (De Geus et al., 2005). The

Figure 4.3: Timeline Childcare Benefits



tax authorities will subsidise the costs parents make. Parents are required to use registered childcare providers, with the involvement of a childcare provider bureau. Parents are jointly and severally liable for debts.

2008 The number of children receiving benefits doubled since the 2005 law (House of Representa-

tives, 2013), for which the main reasons were that previously unpaid grandparents became the official childcare provider and a loophole was found in the law to facilitate this. In 2008, the budget needed 700 million euros more than calculated (House of Representatives, 2009a). Thus, changes in the law followed.

2009 The case of the bureau "De Appelbloesem" is discovered, where the fraud occurred through falsified contracts for which the director got a jail sentence (National Ombudsman, 2010). However, the childcare benefit law states that parents are responsible for the benefit. While the bureau told parents they did not need to pay their contribution, this was not according to the law. Therefore parents had to pay back every euro: their contribution and the whole collected benefit. Juridically, the tax authorities asked for advice from the national lawyer, who replied that it would be possible to collect the whole sum, and then try it in court, intending to create jurisprudence for future cases (Rechtbank Utrecht, 2010).

2010 The law changed and stated that childcare providers needed to fulfil specific requirements (House of Representatives, 2009b), like a diploma of a certain level, be registered in the national registry daycare, a first aid certificate, and fill out an evaluation form safety & risk. Adding to the requirements, the money flow changed to parents paying their contribution and childcare benefit to the childcare provider bureau, after which the bureau transferred the money to the childcare provider. Previously, the money was transferred directly from the parent to the childcare provider. All consequences were still the responsibility of the parents, yet not all bureaus were as professional. Problems occurred when parents could not prove they agreed to pay their contribution.

The House of Representatives asked that the sum of money owed by the parents from "De Appelbloesem" be recovered from the bureau. However, the minister of Social Affairs and Employment replied that this was not possible within the current law and that parents needed to go to court to get their money back (Kamp, 2011). The cabinet (2010-2012) decided that the tax office had to cut down 395 million euros until 2015 (Algemene Rekenkamer, 2013).

2012 Public servants of the tax authorities warn that the earlier tough explanation of the law in court is unwelcoming in this case. However, the first approval of the Ministry of Social Affairs & Employment (Ministry of SAE) is supposedly needed, as they are officially responsible for the childcare benefit law. The ministry is still satisfied with the toughness of the law, as they are worried about an increase in the budget. Ultimately, the Council of State rules in favour of the tax office to continue to enforce the law. The cabinet (2012-2017) decided that the tax office had to cut 126 million euros more (Algemene Rekenkamer, 2013).

2013 The case "Bulgarian fraud" came to light, where people were able to conduct fraud by applying for benefits through incorrect addresses, and the money flowed to Bulgaria. From that moment on, many decision makers were focused on preventing fraud. The management team fraud was set up, creating a new combi team approach facilitators (CAF). In precaution against the "one bank account"-rule, which states that the parents can only receive the benefit on one account with their name, the tax office sent letters to those who needed to send additional information: which bank account does the money need to be transferred, and whose name is on it? Furthermore, the House of Representatives sent a letter in which it was decided that there should be no prior payments when parents are at high fraud risk (145). The tax office starts by judging whether cases are "intentional/huge fault". If the label is assigned, parents have consequences, like not being eligible for a payment arrangement, yet the term is not defined.

The Bulgarian Fraud is only 0,006% of the total transferred benefits.

2014 The "one bank account" rule was enforced and led to slow payment because the benefits were stopped if the information was insufficient. This was a concession in service to limit fraud. The law "Wet aanpak fraude toeslagen en fiscaliteit", the law approach to fraud in benefits and taxation, was unanimously voted in favour of and has been enforced since January. The Council of State states that the law is rapidly enforced, and the changes are impactful.

The Ministry of SAE is working on a new plan for benefits, where childcare benefits are not needed anymore. It prevents them from prioritising the need for a change in the law that parents who did not pay their contribution will only need to pay that amount back.

2015 The National Ombudsman writes a report about the duration of cases and the payment arrangement with parents.

2016 The tax authorities stop collectively cancelling the benefits of parents when there might be misused by the bureau. Half a year later, the council of state judges that it is allowed for the tax authorities to collectively stop transferring benefits when the childcare provider bureau is supposedly acting against the law. This ruling is about a case from 2013.

2017 The council of state judges that it is not allowed for the tax authorities to collectively stop transferring benefits when the childcare provider bureau is allegedly acting against the law. This ruling is about a case from 2014. However, this ruling is one of the only ones favouring the parents. The National Ombudsman writes a report about fair play in the law cases of the parents. A periodic consultation was formed between the head of benefits at the tax office and the head of childcare benefits at the Ministry of SAE.

2018 The plan for a new way to handle benefits from the Ministry of SAE is cancelled.

2019 A hardship clause is implemented in the law's general income-dependent arrangements after consultation between the two ministries.

This timeline eludication shows the importance of taking into account the continuous nature of the system. It is not static and prone to change due to other direct and indirect developments.

4.4.3. Process for parents

The childcare benefits case uses a risk classification model to decide whether a citizen receiving benefits is at risk of being fraudulent or not (Board of Directors Benefits, 2021). Based on the classification, citizens are researched further and often need to deliver additional documents to prove they are not unrightful receiving benefits. The process for the parents in this system is illustrated in figure 4.4

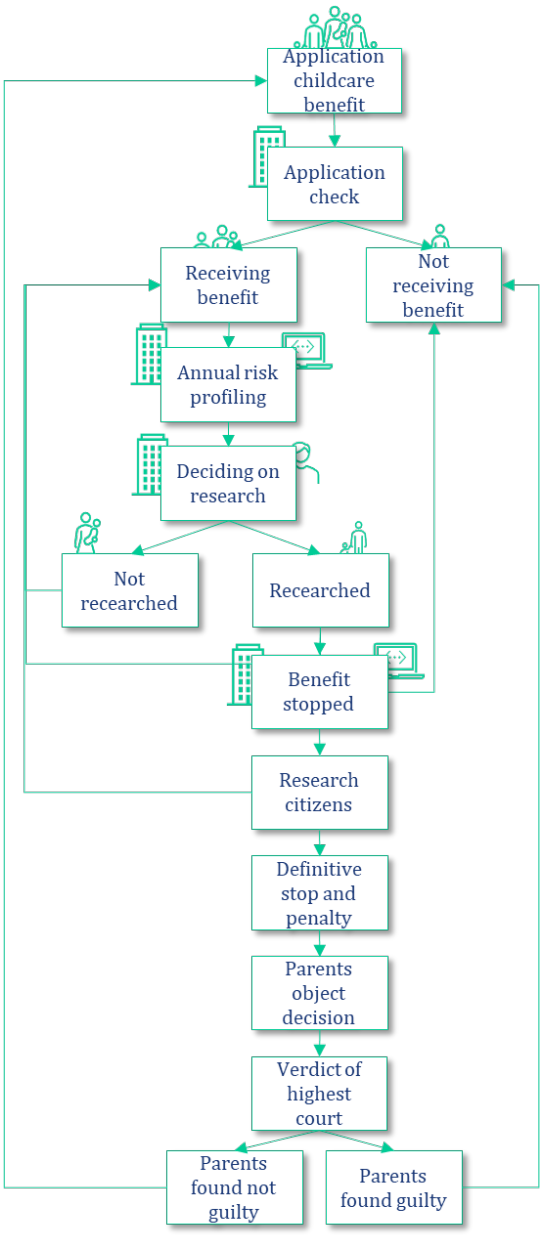


Figure 4.4: Process of receiving childcare benefits and possible consequences

From 2013 (Autoriteit Persoonsgegevens, 2018) up until 2020 (Belastingdienst, n.d.) the tax authorities used a risk model for profiling citizens.

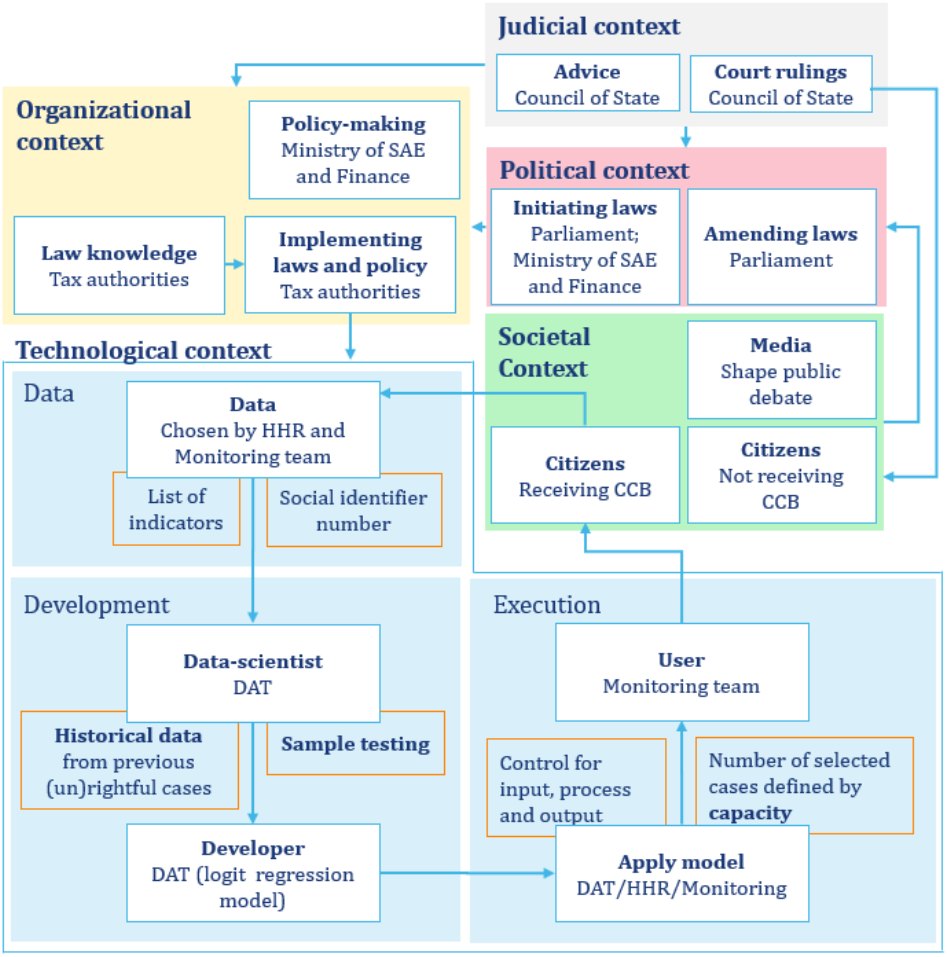
According to CBS (2022b), the number of affected households is 24.125, and the amount of children sums up to 63.120. They counted the out-of-home placements as well, which resulted in 1.675 children moving away from their parents. The total of children for which the benefit is received varies between 0,75 in 2013 and 1,00 million children in 2020 (CBS, 2022c). This means that 6,3-8,4% were negatively

affected, assuming that the individuals are the same yearly.

4.4.4. Systems perspective

The decision-making process can be illustrated as a chain and is portrayed for the childcare benefits case in 4.5, and shows precisely how the different contexts are interdependent. The arrows stand for decisions, e.g. the judicial context influences the organisation, politics and society through interpreting laws, and society influences politics by voting, protesting or other democratic rights, and influencing data by being recorded in it.

Figure 4.5: Process of receiving childcare benefits and possible consequences



The multi-actor system is portrayed through all contexts. Even though in the technological context, the same organisation is present (tax authorities), the context does distinguish different teams or departments that within the organisation can function as actors. Parliament can make and change laws, the Ministry of Social Affairs & Employment is responsible for the childcare benefit law, and the Ministry of Finance is responsible for the general Act on income-related schemes. The tax authorities, which are part of the Ministry of Finance, are responsible for implementing the law by collecting and giving money to citizens. They are also responsible for fraud detection. Additionally, the Council of State acts as the highest court of law and the national advisor.

This landscape is already complicated without considering the influence of an AI model on decision-making. Therefore the technological context is explicitly detailed and included in the illustration. The data is chosen by the monitoring team and HHR, ”handhavingsregie”, freely translated to ”compliance regime”, who oversees the process from data to applying the model (Inspectie Overheidsinformatie en Erfgoed, 2021). The DAT, ”datafundamenten en analytics”, freely translated to data and analytics team,

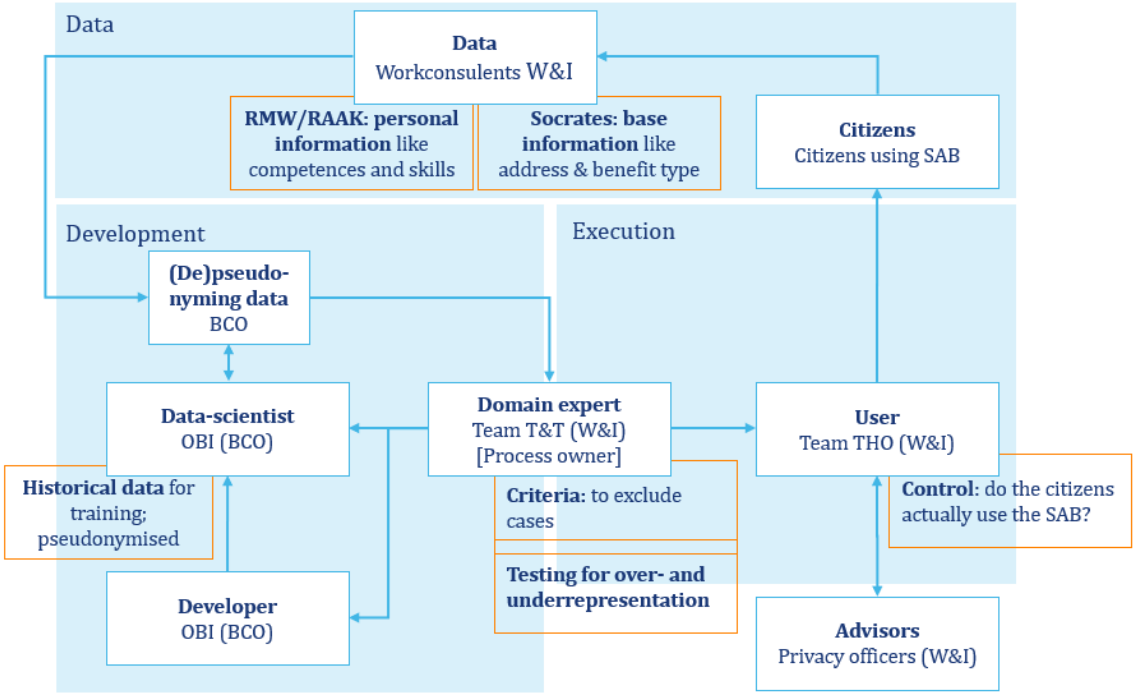
is responsible for the developed algorithm (Autoriteit Persoonsgegevens, 2021).

Furthermore, the illustration notes the character of the stage in the orange rectangles. The data included a list of indicators (Ministerie van Financiën, 2021) and the social identifier number (Autoriteit Persoonsgegevens, 2021). Both were combined to predict the chance of fraud in the development stage, in which historical data and sample testing, combined with the model, were used to create a working model (Autoriteit Persoonsgegevens, 2018). The control for input, process and output is done when applying the model to the new data after the selected cases are defined by capacity, meaning that if they could research 30 parents, they would research those 30 with the highest risk scores. However, this method can eventually be second-guessed as personal data seemed to be important nevertheless (PWC, 2022).

4.5. Social assistance benefits case

The Municipality of Rotterdam (n.d.) uses a risk assessment model to identify irregularities with the Social Assistance Benefit (SAB). It is comparable with the AI model of the tax authorities. The Rotterdam Court of Audit (2021) conducted research into the specifications of the use and risks of the algorithm that are not mitigated. They describe the organisational system illustrated in figure **??**, where the components data, development and execution are distinguished.

Figure 4.6: Organizational Relations



Data for the model is provided by the work consultants of the cluster Work & Income (WI), which exists in two parts. One part is provided by the governmental database Socrates, which entails general information about citizens, like their address and the type of benefit they receive. The other part of the data is gathered from the RMW/RAAK system, which entails more personal information like competencies and skills. How competencies and skills are quantified remains unclear, but they consist of the notes of caseworkers who are in contact with the prospective citizen. The data is then used to train the model by the data scientists from the cluster OBI, involving pseudonymised data for which the cluster Management Support; Governance & Support (BCO) is responsible, as the developers and data scientists are not allowed to have insights into citizens' personal data according to the privacy law GDPR (European Union, 2018). The development and data-science stage are overseen by the domain expert of the team Monitoring & Assessment (T&T; "Toezicht & Toetsing"), in cluster WI. The latter also

sees to the execution of the model by the user, who is part of the team Re-examination (THO; ”Team HerOnderzoeken”) in the cluster WI. Furthermore, the domain expert controls for specific criteria and excludes outputs (citizens) if they were already re-examined in the past two years, they are 65 and older, they are living in an institution, they do not have a registered address. The domain expert also tests for over- and under-representation. How this is explicitly done is unclear Privacy officers of the cluster WI advise the user about privacy concerns. The last control of the user is whether the output (citizens) by the model are receiving benefits. In the end, cases that contained irregularities will be used as input data for the model again as training data. It is unclear whether examined cases that did not lead to any irregularities are also fed to the model as training data.

When the inputted data is researched beforehand on representation for the population and the output data is tested to check certain unintended behaviour, the impact of an AI can be open, requiring knowledge at the implementation part. As seen in the case of Rotterdam in figure **??**, it is clear that certain criteria are checked, as cases are excluded, and tests for over- and under-representation are implemented. It is not clear how this functions precisely, and it is unclear what the exact relationship is between domain expert and user. In the childcare benefit case, an over-representation of single parents and parents with an income smaller than 20K€ (Liem & Nasrullah, 2022) appeared.
Ethical trade-offs are not considered, is the conclusion of the Rotterdam Court of Audit (2021). This is interesting, as this model came into practice only after the childcare benefits scandal. Only privacy is safeguarded to comply with the GDPR. However, it is unclear what measures are explicitly taken by the privacy officers involved.

4.6. Cohesion in explorations

This chapter elaborates on the paramount empirical concepts in this research. Combining the knowledge portrayed, it is possible to examine the relations of the notions and concepts of both explorations. In figure 4.7 the cohesion among both explorations is illustrated, where distinct characteristics have developed.

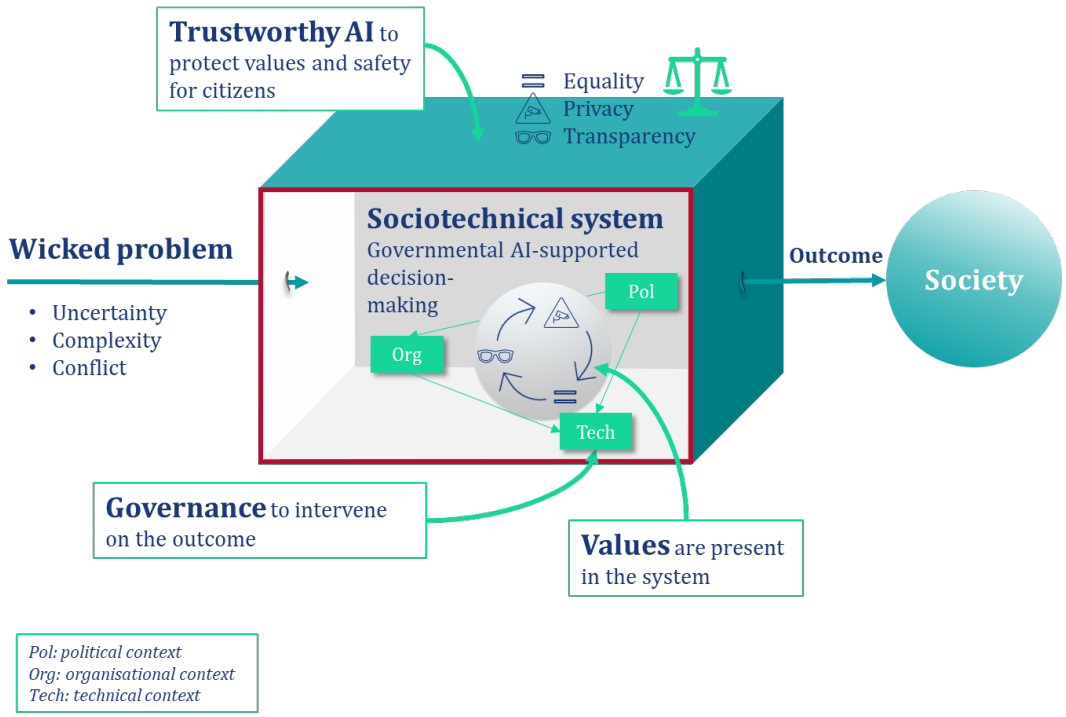
Trustworthy AI is often terminology used by policy perspectives. Therefore, this figure illustrates the change from system safety toward trustworthy AI. In most policy perspectives, trustworthy AI is the notion used to describe a system where safety of citizens is reckoned and the technology can be trusted. The challenges of AI previously defined, come together through the scale portrayed first by the (Prins et al., 2011). These values are reflected in the system as well, portrayed in the cube. There, the political, organisational and technological context are connected as well. Governance is no longer seen as governance on the system, but governance on the outcome. In the cases it is reflected that the outcome is checked by a human decision-maker between the technological context and the fulfilment of the decision.

Values are adapted as factors to understand safety through all components. Privacy, bias, and transparency are all anchored in law. These values are worth pursuing regarding trustworthy or responsible AI. Values can be seen in political, organisational, and technical components, which is crucial when understanding the interdependencies among components. Furthermore, efficacy entails uncertainty and thus cannot be the only factor when considering human lives in a governmental context. Lastly, efficiency is important as resources are public, yet it cannot be seen as the primary value as the governmental context asks for more than efficiency. Values provide insight into the trade-offs made in the system, as a completely transparent, private, and equal system is inherently impossible, which means there are contradictions in the values. For example, if complete transparency is given to the organisation, privacy is probably infringed on. Another example regards the interdependencies between contexts. If complete equality is given in the technological features, privacy must be violated by the final decision-maker to control that this is true. The trade-offs that are made between values in one context influence the decisions that can be made in another context.

The political, organisational, and technological contexts contain decisions that add or reduce the values in the system. The values are interdependent between contexts and summed up the form the balance at which the system can be judged more or less safe internally. Especially feedback loops can be enabled by suitable governance structures. More specifically, the value interdependencies between

the decisions in and between the contexts can make for a more precise intervention and trade-off.

Figure 4.7: Cohesion in explorations



4.7. Conclusion & societal gaps

This chapter gives empirical insight to reflect upon Governmental AI-supported decision-making and its empirical status quo. Insight is given through desk research by analysing countries' approaches and European Union's policies and lastly by laying out details of two Dutch benefits cases, by answering the second sub-question: "What can be learned about governmental AI-supported decision-making empirically?".

First and foremost, this chapter empathises the societal relevance of this research by laying out the details of the Childcare Benefits Case. In this case the political and organisational sphere have changed significantly over ten years time. Crucial to understand is that the harm following the harsh interpretation of the law in the Childcare Benefits Case did occur while respecting the law. Therefore, laws cannot inherently be regarded as safe for the citizens in the system. Another crucial insight this chapter gives is that the citizens affected by the decisions made in the system often have a low social status, have a low income and lack other means, while these citizens may be the ones who need protection most. When detecting criminals, it is difficult to apply the right to remain innocent until proven otherwise, as freezing their accounts is desired around the time of arrest before money. This can lead to high stakes, in which both government and criminal have a lot to gain or lose.

Colliding goals exist for law enforcement if the high stakes when not applying consequences to alleged fraudulent citizens before the crime is proven in combination with the right of citizens to be regarded as innocent before proven otherwise. Linking the low social status to this situation, the societal relevance for research into this topic increases. The system described in this paragraph is major, from law-making to operational implementation, and is not yet researched including the sociotechnical features, i.e. including both the technological and political context. Citizens are protected through the Human-on-the-Loop (HOTL) governance structure, however, whether it functions is not researched empirically.

Governmental AI-supported decision-making bears its challenges. The previous chapter illustrates the challenges surrounding equality, privacy, and transparency. These values are legally embedded

in the Netherlands, the European Union, and other countries, serving to protect citizens from other entities. Especially paramount is the protection of citizens against government, as citizens do not have a choice to be subject to government, distinguishing governmental organisations from other types of organisations. Thus, protection from governmental decisions is embedded through values. As the protection cannot be depending on the current government colour, these values are embedded not solely in regular laws, but embedded in the constitution and international treaties as well which.

Policy is made for creating trustworthy AIs, however, how the transition from policy to operation is executed is a challenge. In the social assistance benefits case several safety measures are implemented to ensure a reasonable outcome of the algorithm. For example, the citizen may not be researched for fraud every year, to be eligible for fraud detection the citizen must receive said benefit, and the citizen must have a registered address. Currently, these measures are focused on preventing accusations toward citizens who cannot conduct fraud or whom cannot be found by government agencies.

The values that protect citizens from government may be infringed in governmental decision-making. As AI literature explicitly names these values as a challenge in AI decision-making, this topic requires more attention. The current literature reflects ideas on how to avoid infringement on the technological or on the social aspect, however disregards a holistic view on both aspects. From an empirical perspective, oversight, human agency, privacy, transparency, non-discrimination, and accountability are posed as the solution to the possible infringements. In this, the Dutch Scientific Council (WRR) sees the public values as a scale that needs balancing. Whether this creates a trustworthy and safe system requires more research.

The multi-actor environment is shaped by the different governmental organisations. Therefore, shaping regulations, processes, and other components of influence are developed through multiple actors. The gap between policy and implementation is a challenge, and how specifically trustworthiness is operated in the system of governmental AI-supported decision-making is unclear. Additionally, the specific requirements for technical systems to be trustworthy or responsible are undefined. Trade-offs are not always explicit in this complex multi-actor system with wicked characteristics. In the childcare benefits case, it remains unclear why certain decisions are made and how actors interpret trade-offs. Trade-offs are either (1) not made explicit or (2) not made public in the implementation. The unknowns in the multi-actor environment results in the incapability to research where, why, and how to enhance the complex sociotechnical system. Therefore, the system requires clearance on who are the stakeholders, what their duties and rights are, and how they execute their tasks.

This chapter concludes that both benefits cases help to understand the system of governmental AI-supported decision-making, however, the Childcare Benefits Case shows that the political arena does not lead to decisions that will not lead to regret later. This adds complexity and uncertainty to the system and the question arises whether citizen's safety can be defined in absolute terms. The technological context is shown in the Social Assistance Benefits Case in which measures are taken to ensure that the correct citizens are captured, starting with the measure that to be an alleged fraudster, one has to receive the benefit. These insights are crucial to design the system artefact.

05 Understanding the system

"What are the crucial system characteristics?"

Content

Takeaways

- The central notions in this research are connected through the system
- The system is about wicked problems, is multi-actor, and has an unknown outcome
- The system can show dynamic behaviour in which system objectives change
- System components are inter-dependent, therefore the citizen's safety might be

5

Understanding the system

This chapter aims to synthesise the previous understandings regarding the rigour cycle (chapter 3) and the relevance cycle (chapter 4). This chapter is established through actors and system analysis and opens with a meta overview combining the scientific and empiric background, after which the system follows in detail. The analysis ought to define the system and its components, boundaries, and characteristics. This chapter discovers the crucial factors and designs a framework which is the foundation for the following chapters. Through the focus points addressed, there is added to the design cycle, depicted in figure 5.1.

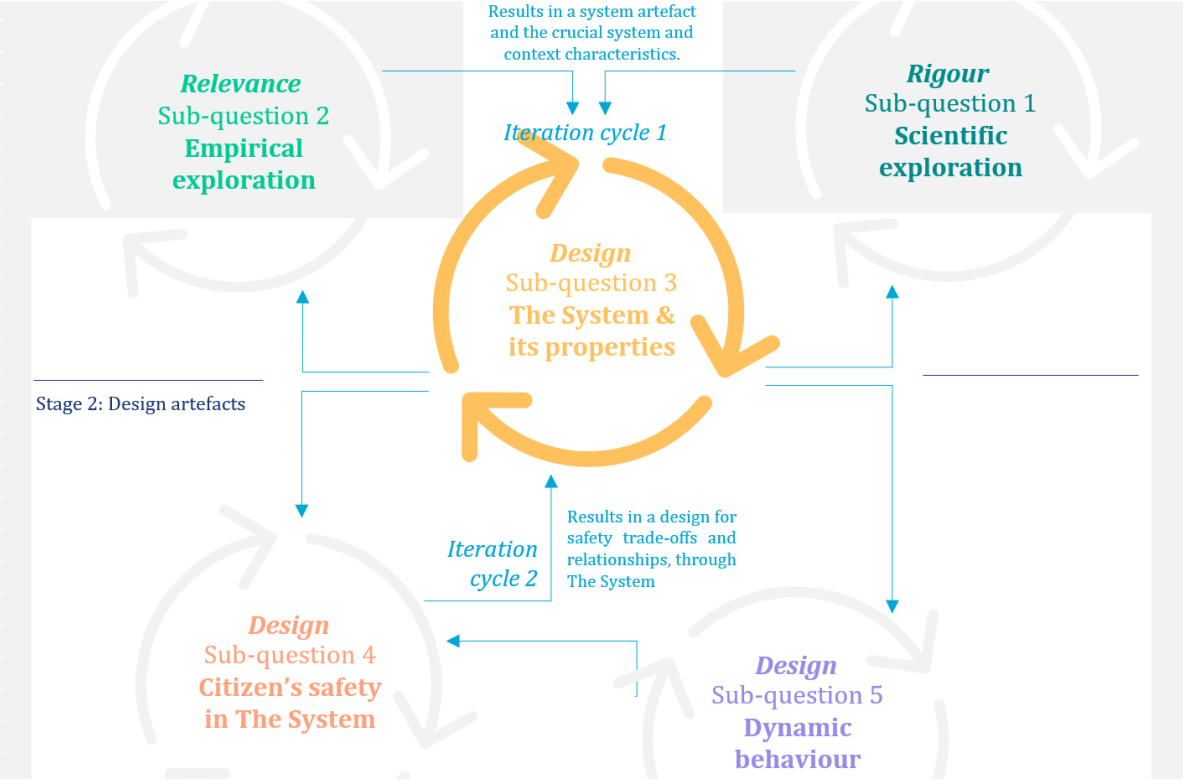


Figure 5.1: Design science cycles relevant for the system, inspired by Hevner (2014, p. 88)

Specifically, the sub-question ”What are the crucial system characteristics” is answered in the following sections—starting with the meta overview that predominantly relates the notions to each other in section 3.6. The structure is followed by zooming into the system box and generalising the contexts in section 5.1 and an elaborated overview of the actor landscape in section 5.2. The following illustrates the defined decision-making chain in which, inter alia, the system and its boundaries are presented in

section 5.3 In section 5.4 discussing the connection of the system with wickedness, linking back to the meta overview at the beginning of this chapter. The conclusion is structured in section 5.5

5.1. Four interdependent contexts

This section aims to give insight into the different contexts of the system, which is constructed as the first step in creating the imminent system. Through different actors involved, the contexts are deduced. Starting with the complex composition of stakeholders in the childcare benefits case, in which the House of Representatives can make and change laws, the Ministry of Social Affairs and Employment is responsible for the childcare benefit law, and the Ministry of Finance is responsible for the General Act on income-related schemes. The tax authorities, part of the Ministry of Finance, are responsible for implementing the law by collecting and giving money to citizens. They are also responsible for fraud detection. This landscape is already complicated without considering the influence of an AI model on decision-making. Additionally, the Council of State acts as the highest court of law and the national advisor. Therefore, the system is categorised into four interdependent contexts. The different contexts of the system are the societal, political, organisational, and technological contexts.

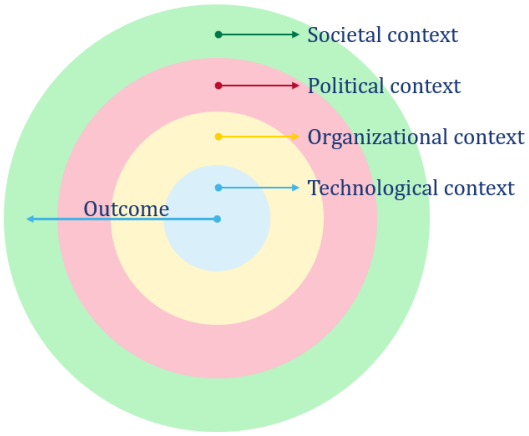


Figure 5.2: Four interdependent contexts

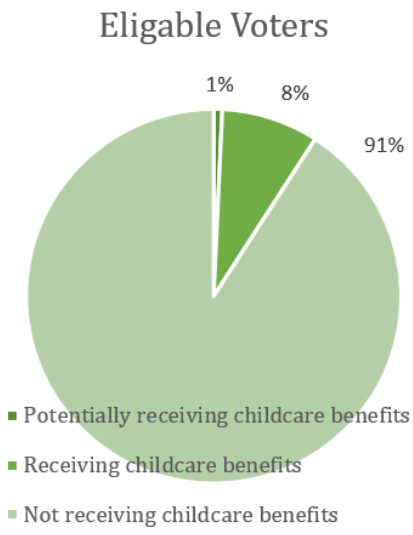


Figure 5.3: Pie-chart of citizens receiving childcare benefits

The societal context consists of the individual citizens. Individuals alone mostly do not have the power to change the system, yet they indirectly can by voting for politicians and political parties. Additionally, citizens can invoke their rights in court against governmental organisations. One crucial actor in this context is the National Ombudsman, who defends the interests of citizens and helps governments improve their services (Nationale Ombudsman, n.d.). On a high level, the system is shaped by society, assuming the government functions correctly.

Addressing the childcare benefits case, there are 13,6 million people with the right to vote (CBS, 2022a), and 0,7 million households receiving childcare benefits (Rijksoverheid, 2020), of which many the same people. Some differences appear due to international people only gaining the right to vote after five legally living in the Netherlands (Rijksoverheid, n.d.-d), resulting in 13,7 million people potentially being eligible for childcare benefits. Translating the number of households to the number of caretakers with the averages calculated by the Central Bureau of Statistics (CBS, n.d.), the number of caretakers in the households totals 1,24 million people. In 5.3, the percentages represent the people certainly receiving childcare benefits with 8%, possibly by 1%. Without any support of the remaining 91%, the 8% could gain 12 seats in parliament, where 150 seats are divided (Constitution, 2018).

In the political context, laws are created by a coalition of politicians, initiated by either politicians from parliament or ministries. The political context consists of the States-General, the parliament and the Senate. The parliament can initiate and change laws, control the administration, and fire ministers. Lastly, as an independent research committee, they can conduct research if deemed necessary (Tweede Kamer, n.d.). The senate approves or rejects the laws voted for in parliament and may control the administration (Tweede Kamer, n.d.). Since 2010, the liberal Dutch People's Party for Freedom and Democracy (VVD) has been the biggest, varying between 31-41/150 seats in parliament (De Neder-

landse Grondwet, n.d.; Parlement, n.d.). Through the political context, the system is shaped further by a legal framework.

The organisational context includes both ministries, executing agencies and inspectorates, which are part of a ministry. In this context, policy is formed by ministries, implemented by executing agencies, and inspectorates control the actions. In this research, the Ministry of Finance and the Ministry of Social Affairs and Employment play a central role. The ministries can initiate laws and form policy (Ministry of Social Affairs and Employment, 2021; Organisatiebesluit Ministerie van Financën 2020, 2022). The tax authorities are part of the Ministry of Finance and collect taxes and awards benefit (Rijksoverheid, n.d.-c). Together the organisations define the solution space and direction for the system through policy, and systems objectives are defined.

The technological context entails the model and data used by parts of the organisation. It is a part of the organisation yet distinguished because it entails different challenges. The relation between the technological context and the others makes for a socio-technical system. It is important to note that the outcome of the models in this context directly influences part of society.

The different layers include different functions. Two types of bodies cannot be placed in the inter-dependent context of the four layers: the judicial powers and inspectorates. The judicial powers are independent and therefore cannot be influenced, yet they influence the system by interpreting laws and regulations, thus creating a one-way interdependence. The inspectorates apply the same construct. NB., the Dutch labour inspectorate, is part of the Ministry of Social Affairs and Employment and, therefore, might be influenced by the priorities or culture of the Ministry. Diving deeper into the relationship between the inspectorates and their ministries is a different organisational approach and is left out of scope.

5.2. Actor landscape

This chapter aims to discover the actor landscape of the childcare benefits case, as in that case, the political and organisational context is rather elaborate yet complex. Additionally, in this case, the actors can be distinguished and categorised for renowned reasons, as this actor playing field is at the central level of government, inter alia defined by the constitution. In contrast, in local governments, the boundaries between actors may not be clear publicly and to the outside world.

The system is multi-actor, as multiple organisations are involved at different decision-making stages. Nevertheless, together, they make the decisions that define the outcome. Politics can be very far from the operational layer in a public system, and developing technical models is often subcontracted as knowledge and means may be lacking within. Subcontractors are left out of the scope of this research because the government bodies would have all the power to impose the system's requirements. The actors with decision-making power are listed in figure 5.4. The Inspectorate Tax Authority, Benefits and

Figure 5.4: Actors and their power and objectives

Actor	Power	Objective
Tax Authorities	Executive, monitor citizens	Grant benefits, prevent and counter fraud
Ministry of Finance	Legislative, policy	Financial healthy government
Ministry of Social Affairs and Employment	Legislative, policy	Social healthy government
Council of State	Advise, judicial	Functioning rule of law
National Parliament	Legislative, monitor	Shape a good government
Senate	Legislative, monitor	Shape good legislation
"Ombudsman"	Partly judicial	Handle complaints from citizens about government
Dutch Labour Inspectorate	Monitoring Ministry of Social Affairs and	Track down large-scale fraud operations
Inspectorate Tax Athority, Benefits and Customs	*Monitoring Ministry of Finance and Tax Authority*	*Independent and transparent monitoring, recover trust*

Customs are shown in italics, as they were only recently initiated (Official Gazette 2022-4749, 2022).

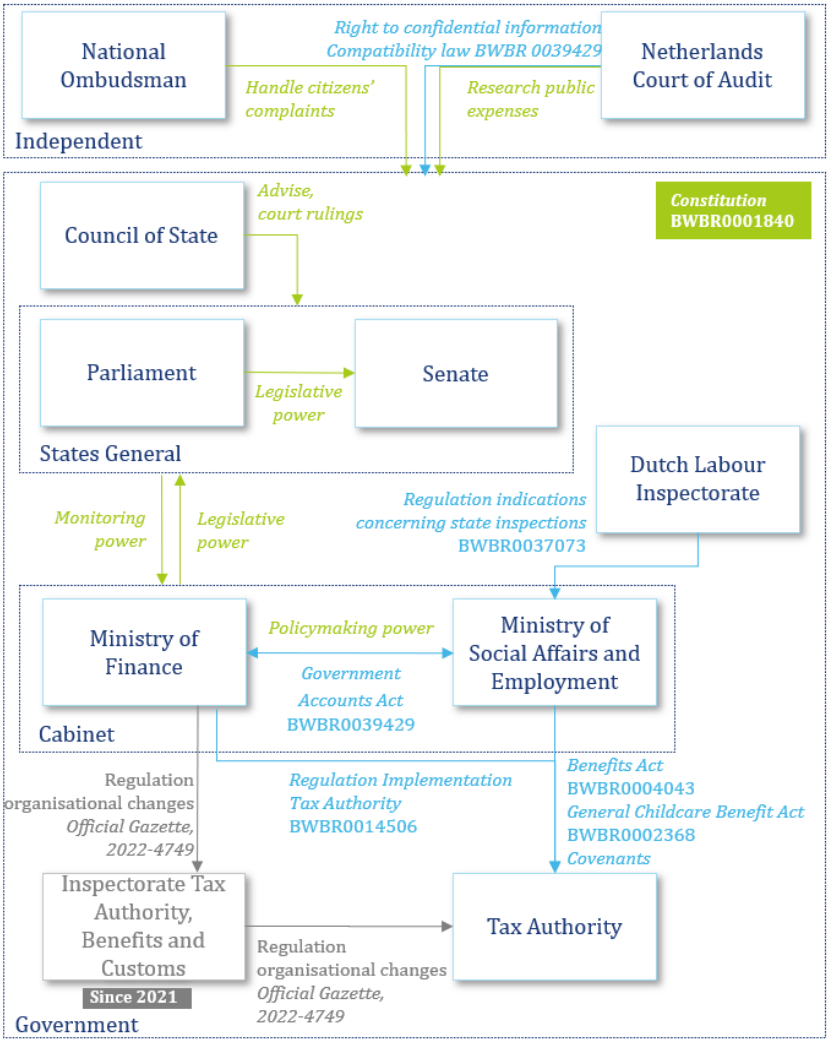
In this research, it is assumed that the actors involved are cooperative: all should agree that money should be received by those who have the right to receive it. The only exemption is the citizens conducting fraud scattered throughout the population in society. However, their influence as an actor in

the system is limited because they are individuals. The other actors in the system all want to obtain the same outcome. Nevertheless, their beliefs on how to get there, their power, and their exact objectives and roles may differ. Small teams and individuals within the different contexts are left out in this section, as they are part of the high-level objectives and power.

The relevant actors in the system are mainly governmental and have legally binding relations. For actor analysis, it is helpful to define a problem of a particular problem owner, who can intervene in the system (Enserink et al., 2010). In the childcare benefits case, the responsibility is divided. The Ministry of Social Affairs and Employment is responsible for the benefits law. The Tax Authority is responsible for monitoring and is part of the Ministry of Finance, which is responsible for implementing the law (Benefits Act, 2022; General Childcare Benefit Act, 2022; Regulation implementation Tax Authority, 2022). Therefore, both ministries and tax authorities are central nodes in the actor landscape.

This claim is supported by the standard chart of actors, displayed in figure 5.5, in which the formal relations between actors are shown concerning benefits. To keep the systems approach, one actor cannot be chosen above another. Therefore, the actors' analysis considers the perspective of both ministries and tax authorities. This is desirable because the main high-level goals of the different governmental organisations are similar.

Figure 5.5: Formal chart of actors



Adding to the aforementioned section, the actor landscape also knows regulatory bodies. The inspectorates are responsible for monitoring the governmental bodies, whereas the Council of State functions as an advisor and a court. The Inspectorate Tax Authority, Benefits and Customs is shown in italics, as this body only came into existence in 2022 (Official Gazette 2022-4749, 2022). This is also

why the formal chart in figure 5.5 the organisation and its regulations are displayed in grey. Furthermore, the green arrows refer to the Constitution (Constitution, 2018).

Both ministries are central nodes, supported by the formal chart of actors, displayed in figure 5.5, in which the formal relations between actors are shown concerning benefits. The Tax Authority is the only actor in this chart with only incoming arrows, meaning that they are the last node before they are met with citizens in the case of childcare benefits. Thus, it is the executing agency and is in line with what is expected in a formal chart. To keep a holistic approach, one actor cannot be chosen above another as the most important. Therefore, the actors' analysis considers the perspective of both ministries and tax authorities. This is desirable because the main high-level goals of the different governmental organisations are similar, and they are all in one context, the organisational layer, as discussed in the aforementioned section.

At the political level, politicians are spending their time on all challenges in need of legislation, their part-politics, and the next election. These factors can detract from AI decision-making at executive organisations. At the policy level, they have to translate laws to the policy that fits, and laws leaving room for interpretation, policymakers can always fill in policy in what fits best. At the executive stage, the actor needs to have the skills to understand the policy and know how to translate this for use in a model requiring technical knowledge. Here is defined how the model is used and developed, including making judgements about whether to include competencies and skills of citizens or not. The more clear the guidelines are in the processes, the more clarity is given to the developer on the specifications of the model design. It becomes complex when other actors manage the data and decide what data could be useful, how it is processed and how missing values are handled. Concluding, there are actors involved in every stage of the system, and they all have a certain space in which they make decisions. Partly because of the considering objectives of the system, rightfully detect fraud, and the objectives actors might have for themselves.

In conclusion, the formal chart of actors, figure 5.5, shows the interdependencies between law-making and law-executing. The tax authority is influenced by two ministries, creating the possibility for conflict when both ministries have different perspectives on how the tax authority should execute their policy. Additionally, the lower part of the chart is influenced chiefly by laws made by the states-general, whereas the constitution is the most robust form of law in the Netherlands. In this chart, it is seen that through law as well, feedback mechanisms are minimised.

5.3. The system

The goal of this section is to convey the characteristics of the decision-making system. As with the framework of Rasmussen (1997) in section 3.3 on socio-technical systems, this section also presents a chain. It is presented in section 5.3.1, whereafter the outcome characteristics are portrayed in section 5.3.2. The following section explains the present uncertainty in section 5.3.3 and information asymmetry in section 5.3.4. Lastly, this section elaborates on the factor of time in the system in section 5.3.5.

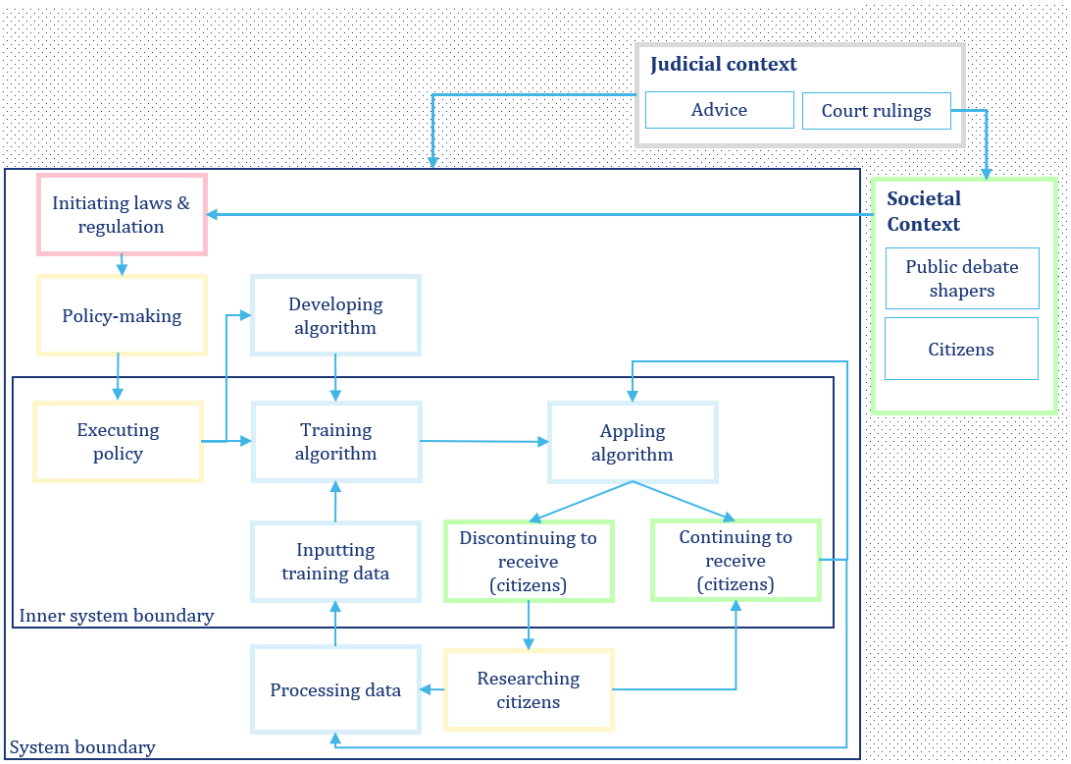
5.3.1. Synthesised systems overview

This section aims to clarify the system and its boundaries from the previously obtained knowledge. The system is portrayed in figure 5.6. The colours of the different stages, the action-based rectangles, are consistent with the four distinct contexts explained in the second section of this chapter, 5.1. Two system boundaries are depicted in black: the system boundary and the inner system boundary. The inner system boundary denotes the relative short-term decisions, and the outer system boundary the more long-term decisions. Central in the system is decisions. In every stage, decisions influence other stages, illustrated with an arrow.

The systems description starts with the external factors, defined by the judicial and societal context. The arrows stand for the influence of decisions on the subsequent decisions, e.g. the judicial context influences the organisation, politics and society through interpreting laws and advising about the system's functioning. The judicial context influences society with court rulings and, thus, jurisprudence. Society influences politics by voting, shaping democratic power, protesting or other democratic rights, and influencing data by being recorded. The external environment is synthesised from empirical exploration and the principles of constitutional democracy.

Next is the political context, where laws and regulations are initiated. The law-making process is

Figure 5.6: Different contexts connected for the childcare benefits case



out of scope for this system, as only the final decisions on laws and regulations directly impact how the system is shaped. After laws and regulations have passed, the policy is made. Both organisations can be the same, even though the goal of their decisions differ and therefore should be depicted as different stages, consistent with the division of power, Trias Politica by Montesquieu (Von Bóné, 2019). The decisions made at the policy-making stage directly influence the objectives and manner of the executing policy stage.

The executing policy stage lies inside the inner system boundary and is done in the short term. The policy is executed, for example, every year, as is the case with both cases from the empirical exploration. The decisions made in this stage are organisational, as it is part of the policy process and is inherently operational, therefore not prone to politics. The technological context has its first two stages: developing and training the algorithm.

Developing the algorithm again lies outside the inner system boundary. The reason is that not every time the cycle is executed, a different algorithm is developed and is placed between executing policy and training the algorithm because only after the execution is started the solution (algorithm) can be developed. The initial algorithm needs to be developed before it can be trained.

Training the algorithm means that in combination with decisions from executing policy, the initially developed algorithm and the data for training a metric for categorising the data is developed. In the case of predictive analysis, the subject in childcare and social assistance benefits training, the algorithm results in a scorecard where specific values are determined to have a certain weight to calculate the risk score. Thereafter, the algorithm is applied in the next stage, using the results of the decisions made in training the algorithm and the citizens with whom a transaction is made; in the benefits cases, the transaction is the received benefit. A feedback loop has been initiated: the citizens receiving will serve again as input the next time the algorithm is applied. Another possibility after the algorithm is applied is that one stops receiving, meaning that the transaction is stopped immediately, and is derived from the childcare benefits case where preliminary stops were allowed and thus executed. Both the continuing and discontinuing stage belong to the societal context as they consist of the decisions made directly by citizens. However, they can only be with the citizens with a transactional relationship with the government. Important to note is that humans make these decisions. Automated decision-making when profiling is restricted (General Data Protection Regulation (GDPR), 2016) but in both empiric

cases AI-supported decision-making is conducted. An employee intervenes between a person being categorised by the algorithm and the judiciary consequences (Autoriteit Persoonsgegevens, 2018)—concluding on the last stages that are part of the short-term decision-making system.

After discontinuing one's transaction, one is researched by the government to obtain knowledge of when one did rightfully transact. Important to note is that it can be decided that one was indeed rightful, flowing back to the continuing to receive, creating another feedback loop toward applying the algorithm again, possibly again being discontinued to again being researched. As this is an intensive process for the researched citizens, the social assistance case implemented a measure that a citizen can only be researched every three years.

Following the research, the data gathered on those who received wrongfully are processed to be able to use as input data in the training algorithm. The inputted data saves up the cases and patterns of those who were wrong to capture the patterns into the metric for categorising the data. With this, the feedback loop is closed.

In conclusion, the decision-making system is not only a chain but includes functioning feedback loops, introduced explicitly through the technological context considering this topic and the need for human decision-making. Therefore this framework is similar to that of Rasmussen (1997), explained in section 3.3, yet adds a better abstract representation for this topic. The need for human decision-making is the first difference with the framework of socio-technical systems, as there the outcome of the technological layer can be implemented, which can be true, for example, for factories. However, other approaches are needed in a system where the problem is not linear, such as in figure 5.6.

5.3.2. Outcome of the system

Decisions in the system are made consecutively, from a call for change to laws, policy, processes, and eventually, an outcome: the final decision. A chain of decisions is made throughout every stage. The technical approach is a new link in the grid, where there is also space for decision-making. Important to note is that the call for change is present. After all, there is already a system in place as current systems cover (almost) all: therefore, in reality, the decisions that are made are the ones that differ from the status quo, resulting in interdependencies between stages, where for example, laws are practically unable to make rigorous changes to the whole system, as to deviate from the status quo because the organisational layer is not able to catch up. However, in the end, at the outcome stage, it is possible to assess how the system works.

The outcome is defined at the end of the decision-making cycle when the decision is making its way into society, i.e. the alleged fraudster is researched and serves as input for the next run. Therefore, understanding the outcome is crucial for system insight. In the case of risk models for fraud detection, the outcome, in reality, is binary: either one conducted fraud or one did not. The severity of fraud can be distinguished into categories, and the definition of fraud needs to be explicit. Is it fraud if one did not have the intention of fraud, or is it fraud when one makes a simple mistake? In the childcare benefit case, one was conducting fraud even when minor mistakes occurred, yet this changed over time but still lacks a specific definition. Vagueness in the decision-making chain is a threat to a well-considered decision. It is already present in the organisational and political context, which adds complexity to the technical context.

Technology's crucial effect is explicit definitions, and measurable Key Performance Indicators (KPIs) are needed to make functioning software. When definitions do not exist explicitly, it can have far-reached consequences, e.g. unfounded allegations lead to severe debts, which lead to the out-of-home-placement of a child or children.

Adding to the complexity of the outcome is the unknown efficacy rate. Some boundaries can be set through clear definitions, e.g. one conducts fraud when one receives childcare benefits yet does not have any children (Rijksoverheid, n.d.-b). However, the allegations can become greyer if one is allegedly a fraudster because they do not have a significant income or are a single-parent household. In the childcare benefits case, the tax authorities had the right to initial stop benefits transactions. When someone is allegedly conducting fraud, it is complicated to judge the truthfulness without causing inconveniences to that citizen. Additionally, if that citizen objects, the court rulings can take much more time before one is found guilty. Especially if one of the parties does not agree with the trial result, there are different stages of objections that can be carried out in court.

The model is fed with supposed non-fraud cases in the childcare benefit case to train the model to categorise low-risk cases. The training is complicated as it is difficult to prove one's innocence,

especially when one is not researched. Therefore certain assumptions are made to gain enough ”low risk” cases to train the model, which makes the outcome uncertain, as the best criminals are the ones they would never suspect. Adding to this, if one knew the exact requirements to end up on the high-risk pile, one could shape the fraud to explicitly not fill those requirements, which are always crucial in detecting fraud, whether with technical or manual means.

It results in the inability of an immediate judgement of truth. Therefore, the inability to direct continuous feedback to the system based on the outcome.

5.3.3. Uncertainty

The technical system has three defined uncertainty types: the uncertainty of righteousness of the ”low-risk” and the ”high risk” categorisation and the outcome. W. E. Walker et al., 2013 writes about different levels of uncertainty.

The outcome of the model is dependent on the uncertainty of the data. Nevertheless, it can be relatively certain only based on the algorithm, which can be categorised as a level three uncertainty, as the model can create several ways to predict and, therefore, several system models. The lack of proven causality in the social domain creates difficulty in incorporating the proper mechanisms, seen in the childcare benefit case through the variables included in the risk model being decided upon based on tacit knowledge (Board of Directors Benefits, 2021). Eventually, the AI chooses a definite model, which creates the feeling of having a level one uncertainty, which results in a single estimate of the weights.

The fraud cases in the data used for training are relatively certain yet never wholly. It might be that incorrect cases slip in, which is recently illustrated by the childcare benefits case, in which people were incorrectly ranked as high risk, which was used for training the model later on. Because the risk categorisation does not automatically lead to the decision that one is conducting fraud, the interference of the research, later on, is essential as well and influences the training data. In this research, undesirable influences may occur, like institutional racism (Van Rij, 2022). It makes the proven fraud cases still uncertain, yet not as uncertain as the outcome and leads to a level one uncertainty, in which one is not able or willing to measure the uncertainty (W. E. Walker et al., 2013).

The non-fraud cases in the data used for training have a higher uncertainty than the fraud cases. The cases are identified not only if it is proven that they are not fraudulent but also if they fulfil specific requirements, like a lack of mutations (Board of Directors Benefits, 2021). It would create very skilful fraudsters not to get recognised by the model and even to be categorised as low risk. When cases with specific characteristics are automatically used as no-fraud cases for the model’s training, there is deep uncertainty about which cases are not rightfully categorised. Nevertheless, ideas about alternatives can be identified, which is a level four uncertainty.

The uncertainty adds complexity to the unknown outcome. In the case of fraud detection, it is clear that there is no room for error, demonstrated by the notion of wickedness in chapter 3, which leads to a difficult way of handling uncertainty. If the outcome were known, it would be easier to measure the system. Nevertheless, how many cases are accepted to be wrong arises. If the answer is none, uncertainty should be mitigated completely. If it is impossible to measure the system by the outcome, then another way of measuring should be used to compare different systems and investigate interventions.

5.3.4. Information asymmetry

Inherent in an extensive decision-making chain is the occurrence of information asymmetry. Information asymmetry starts with the delegation of tasks from politics to executing agencies and creates a principal-agent relationship, where politics acts as the principal and the agency as the specialised agent for a better quality of policy execution and efficiency (van Thiel & Yesilkagit, 2011). The principal-agent theory can, to a certain extent, be extended to the relationship of Parliament (principal) and ministries (agents) as well (Saalfeld, 2000), or eligible voters (principal) and politicians (agents), however in the latter a lack of control is noted (Canes-Wrone et al., 2001). The principal-agent relationship between the Ministry of Social Affairs and Employment and Finance and the tax authorities is acknowledged by the establishment of the inspectorate to monitor the tax authorities.

Information asymmetry was exposed by the childcare benefit affair when thousands of parliamentary questions were asked during and after (Frederik, 2021b), where in 2008 they were more related to improper use of childcare benefits and in 2020 more about the wrongdoings of the tax authorities and ministries (Tweede Kamer, 2022). Asking questions is one of the instruments of Parliament able to

close the information gap between the bodies. In this case, the principal-agent relationships are not based on moral hazards or adverse selection because the high-level objectives align with each other for all actors (Gailmard, 2012).

However, when the tax authorities' budget was cut to incentivise fraud detection, they might not have agreed to execute their common objectives. The Ministry of Finance used this instrument. They tried to incentivy fraud detection, which impacted the tax authorities' budget and might have reduced the trust between the principal and agent. Additionally, the tax authorities failed to successfully blow the whistle on the harsh modes of the law towards both ministries or other actors.

One illustrative example of information asymmetry is contact with citizens. When a problem occurs in the benefit, and more information is needed, or the tax authorities stop payments, the first contact for citizens is the tax authorities. They were the first to know the implications of the law, as they are at the operating end. Nevertheless, the Ministry of Social Affairs and Employment could not have directly known the problems as they were not in contact with citizens. Another example is the information or more specific knowledge the different stakeholders have about laws. The involved laws and rulings were difficult to understand. In hindsight, the strict outcomes could have been avoided by either the Ministry of Finance, the Ministry of Social Affairs and Employment, or Parliament. Whether the stakeholders were conscious of this aspect or did not want to take responsibility can be debated. However, assuming that government employees and representatives have the best interest at the heart of their citizens, it is plausible that they did not know.

The information asymmetry is essential as the decision-chain most definitely is influenced by it, yet how precisely it is unclear at this time, but might relate to the outcome of the system.

5.3.5. Time as a factor

Over time system components change. The call from society, the political agenda, organisations, and the solution in the system (technology). It takes time to reveal the efficacy of the outcome, and ethical considerations change over time. Time adds complexity and calls for a continuous approach instead of a static one.

Over time objectives, goals, and even involved organisations change. Whereas in 2005, the objectives were to reduce costs and increase the quality of childcare, around 2009, the focus came to lay on the childcare provider bureaus; in 2013, the objective was to prevent fraudulent individuals. In 2020, the objective changed to explicitly accommodating and recuperating affected parents. A year later, in 2021, employees of the tax authorities warned of overcompensation (Splinter-van Kan & Hol, 2021), and someone who, in 2013, was seen as a fraudster and explicitly judged for it in court is now seen as a victim (Frederik, 2021a).

Thus, time is a relevant factor, and the case cannot be seen outside that period's societal, political and organisational context. How trade-offs are made can differ in a year. It emphasises the need for explicit trade-offs and decisions made in all layers to be able to research the system correctly. The values of detecting fraud and serving citizens follow a sinus-like form, in which, over time, detecting fraud or serving citizens is of utmost importance. Because of unforeseen implementation costs of the law in 2008 and the uproar about fraud in 2013, the interest in servitude lessened, and the interest in fraud detection increased. These are not necessarily opposites, but operating from a state of mind that every benefit-receiving citizen might be a fraudster is different from wanting to make sure all citizens can get by.

5.4. Wickedness in the system

This section aims to connect the previously defined term wickedness with the designed system framework in figure 5.6. This is done in two ways. On the one hand, the system is debated in terms of the characteristics defined by Rittel and Webber (1973), aforementioned in section 3.4.1. On the other hand, the wicked characteristics of complexity, conflict and uncertainty are reflected in light of this system.

If citizens are calculated to be high-risk, every hypothesis about what fraudulent transfers look like depends on the idea of how to solve it. One of the characteristics of fraud is deception (Blakeborough & Giro Correia, n.d.): the tax authorities are unaware of the precise size of the fraud. It will never be known if and whether there would be no fraud at one point in time. So, it would never be known when to stop, as long as benefits are transferred. Because of the unknown outcome in the short term, it is only

possible to do better or worse. Using models to detect fraud can be better or worse than depending on tacit knowledge. The approach to detecting fraud can influence fraudsters and make them change their tactics (Blakeborough & Giro Correia, n.d.), which adds to the inability of a 'true' solution.

In the long term, it is possible to examine the amount of fraud caught, but it would never be possible to know how much percentage of the fraud is caught. It is complicated to determine whether one did not commit fraud, and in the childcare benefit case, this is solved by asking for more information. However, it may be that one does not have the best administration at home and thus cannot hand over the proper documents. The determinants of whether one is a fraudster are ambiguous. Whether to include nationality, living situation, income or the skills & competencies of citizens, political preference, the weather, shoe size, ability to speak or other variables to test if one is potentially fraudulent is undefined.

It is not possible to test the approach against fraud without influencing its environment. Therefore, it is impossible to test a solution without influencing the environment and changing it, which makes testing the following solution happen under different circumstances. Additionally, the system in which fraud detection is incorporated is essentially unique. Therefore, the way fraud detection is embedded is unique. No governments operate exactly alike, although there are more and less similar systems. Generally, governments that carry out redistribution laws are more alike than those that do not. However, copying and pasting it into a different system would never be possible.

At the base of fraud detection, there is a system of benefits. It is a complex social system that should take care of the redistribution of money, but how this is undoubtedly well designed is unclear. In this case, again, there are better and worse solutions. At the root of this problem, there is unfair distribution of means in the first place. How to tackle this again is not in terms of "good" but better or worse. Capitalism or communism? There is no "good" answer but better and worse solutions.

For fraud detection, should one focus on being a serviceable government towards citizens or controlling to ensure fraud minimisation? Should the system be waterproof, and how can supervision be organised? To explain fraud detection is of influence to the potential set of solutions.

In many cases, governments are not allowed to be wrong. It is about people's lives and their well-being, and government should not harm them. Additionally, citizens do not have a choice not to do business with their government. Mistakes from (democratic) governments are generally condemned, as with fraud detection.

As previously mentioned, actors in the decision-making chain have a high-level agreement on benefits, as they are prone to be conflict-averse (Aanwijzingen voor de regelgeving, 1992; Algemene wet bestuursrecht, 1992; Constitution, 2018). However, conflict can emerge when there is no agreement on the lower-level establishment. In the childcare benefits case, different system characteristics increase the location conflict that might emerge.

The decision-chain is extensive, with at least six bodies interfering with benefits (5.5, which influences the bureaucratic characteristics, complicating communication when the tax authorities have to be accountable or change ways with the prior bodies in the decision-chain. Complementing this argument is the absence of an ultimate test proving the challenge to be solved and the impossibility of trial and error in the childcare benefits case.

This leads to the lack of publicly known explicit and measurable KPIs, therefore raising the question of whether there were any non-related to money because the exemption of this is the finances. Finances were cut to the extent that the caught fraud should make up for it; if not, it had to come out of the tax authorities' pockets. Through measures alike, the common normative perspective is narrowed for the involved actors, especially the executive agency in this context. They are essentially helpless against measures like this because they do not hold any regulatory power.

Conflict emerged about responsibility as well. With the Ministry of Finance responsible for the tax authorities, yet the Ministry of Social Affairs and Employment responsible for the benefits law, responsibility for both became no responsibility to any. It can be seen as a form of conflict. Lastly, the culture might have contributed to a conflict-averse environment, which resulted in a lack of tolerance for failure.

Uncertainty is present in many factors of the system. The most influential is the uncertainty about the outcome because the outcome can only be adequately evaluated over time, direct feedback is lacking, and the absolute number of fraudsters is unknown. This type of model uncertainty will be somewhere between levels two and four, depending on the model performance indicator. The error ranges of the model can be calculated to estimate the accuracy. However, the model cannot calculate what essential

factors are missing or what crucial external factors are not considered.

Additionally, uncertainty exists about how the individual decisions of and in the involved bodies add up to the process and outcome and increases with inevitable bureaucratic layers. It might not be of importance when all goes well, yet it is when it does not. Then it increases conflict between and within bodies. This process uncertainty exists less in the political layer, as most documents are publicly accessible and the evolution of laws is carefully documented. However, reducing content-related uncertainty is a challenge for all contexts. Reducing uncertainty in both the technical and the organisational context requires resources in terms of knowledge, time and a budget, especially when solutions are not general but tailored by nature.

Complexity is added through distinct rationales. The lack of causal relations when conducting fraud, the many bodies shaping the system, the information asymmetry, the conflicts between actors, and the uncertainty of both process and outcome all add to the system's complexity. There is no final solution or correct answer in this system. However, more simple is the problem in the case of childcare benefits. It must be stopped if a citizen receiving the childcare benefit is committing fraud. However, much complexity exists in the definition of fraud and the dependency on the zeitgeist, the short-term unknown outcome of the model and research, the absolute number of fraudsters in the population, and the information asymmetry between and within bodies of the multi-actor and multidisciplinary characteristic. For example, manually checking the results and drawing a policy-based line to the maximum allowed risk score before research can increase complexity as the employee makes individual decisions about the citizens up for research, which can alternate the further decisions made by those responsible for executing the research.

In conclusion, the wicked characteristics that object to the problem reflected by this system are present. The presented system in this chapter is inter alia built on the wickedness in the system, and because of it, that is the reason why human decision-making is essential. This section demonstrates that the problem with categorising citizens as high or low risk is ambiguous and wicked, which is increased as the outcome is a high-level uncertain, even with the potential to go as far as deep uncertainty, depending on the specific problem at hand. Even though conflicts are avoided, the elaborate decision-making stages create a challenging environment where contradictory decisions are easily obtained. Uncertainty is present at all decision-making stages and primarily in between stages. The interdependencies are unknown; thus, how one's decisions are affected by the preceding decisions is unknown. What is clear is that the system allows for the stacking of decisions to which a lock-in most likely poses a threat. In short, the wicked problem is crucial for the decision-making systems design presented in this chapter.

5.4.1. Safety in The System

What a safe system is in the context of The System may be interpreted differently due to certain characteristics. The System's outcome is unknown, hence it is not possible to check whether the system is optimally used. Discontinuing the benefits of those who are committing fraud and continuing the benefits of those who are in need of social assistance is the main goal of the system. Empirically, finding this balance for a continuous time is not possible. Therefore, the question arises when the system behaves optimally. Citizen's safety is part of that optimal behaviour.

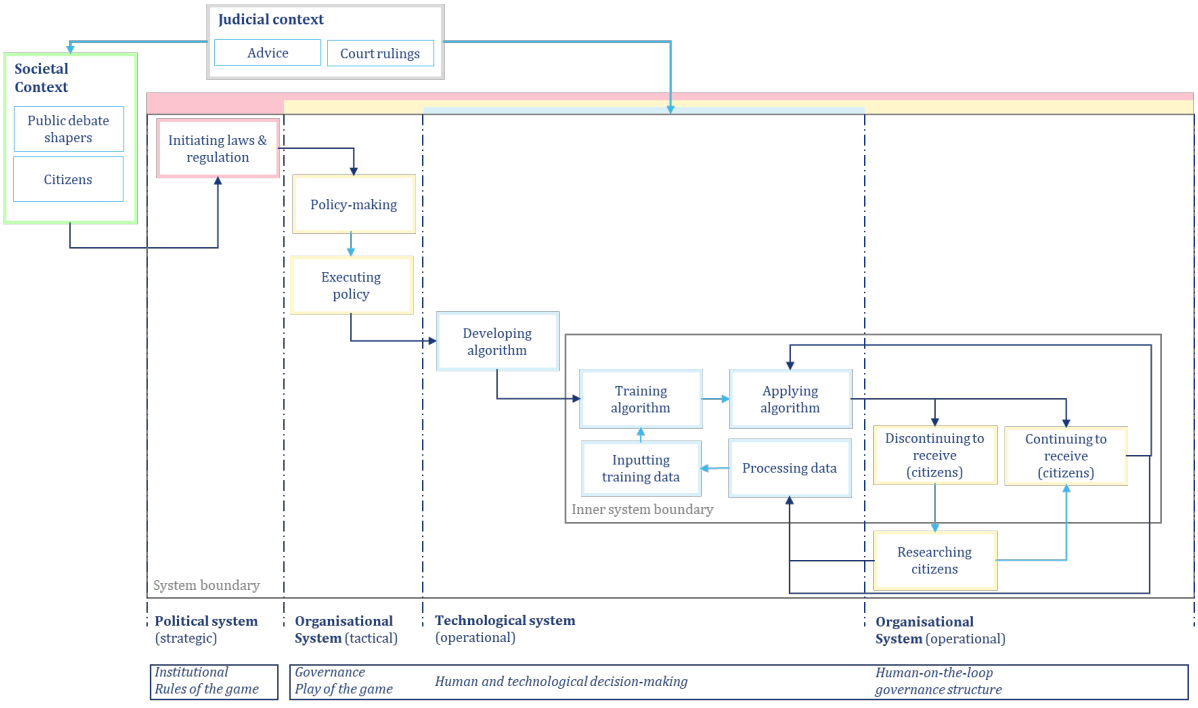
To fit citizen's safety in sociotechnical systems to governmental AI-supported decision-making, the subsystems can be considered, yet with the reason of emergent properties require using the same safety aim. The subsystems in of The System can be derived from the contexts they are placed in: the political arena for amending and initiating laws and regulations on the strategic level, the organisational arena in which policy is decided on the tactical level, and finally the operational system. In the operational system, the technological system is apparent, and the last organisational stages on making the last decision can also be categorised as operational. The system is restructured to fit the different levels in figure 5.7.

The definition of what safety means for the system is dependent on the level in which it exists. On the operational level, covering management, staff and work, compliant with 3.2, safety can be described through measures illustrated in the social assistance case, e.g. not researching citizens every year and checking whether the citizens in the output of the technological system actually receive the benefit. These measures add to the goal of not burdening innocent citizens, i.e. if they were proven innocent

lately they do not have to prove that again soon, and checking whether the decision is logical with the goal, i.e. catching criminals who receive the benefit. The overall aim in this are coherent with the System Objectives, however, safety is an emergent property and need to be assessed relative to the complete system. By checking whether the citizens in the outcome of the algorithm receive benefits, that does not mean they are guilty. On the tactical level, standards are implemented at the level of ministries and executive organisation, monitored by inspectorates. The standards implemented at this level for the system are not apparent when referring to the case study on benefits. However, when the system objective ”detecting fraud” is applied, the standards within ministries and executive agency became to research anyone who may be suspected of fraud. In this, all small mistakes were defined as fraudulent, and became the standard in the political debate as well. When a parent did not pay the own contribution required to be eligible for benefits, the whole sum of received benefits that year required reimbursement. Politically, in parliament and ministries, this was seen as harsh yet needed. In 2015, a motion to collect reimbursement only of the amount of money required for the own contribution did not pass parliament (Tweede Kamer, 2022). On this level, the opinion on how to achieve a safe system depend on the political opinion and definition of what safety means. On one hand, safety can mean social safety, by providing public funded benefits for anyone with a low income. On the other hand, safety can mean financial safety for public spending only funding those who deserve it. The definition of those who deserve it can differ among actors as well. Therefore, the regard and implemented means for safety depend on the System Objectives defined for The System.

System safety in this system can be helpful to understand the human-machine interventions that can be undertaken, placed in the organisational or technological context. In figure 5.7 the system is displayed in a hierarchical order. However, system safety is not useful when safety is ambiguous and depending on law and regulations formed in the political context. Unfortunately, due to the involvement of the political context and therefore the wicked characteristics accompanying this complex problem, system safety cannot yet be helpful. Therefore, this terminology is left and the focus will lay upon citizen’s safety of the citizens in the system. This term covers the protection of citizens again. From a policy perspective, human oversight is argued to allow for securing these rights. However, in science it is argued that human oversight cannot provide the functionality required to perform that task sublime.

Figure 5.7: The system displayed by strategic, operational, and tactical levels



5.5. Conclusion

This chapter discovers the crucial system characteristics through system and actor analysis. The interdependent context is discovered to position the decision-making stages and rightfully explore the complex actor arena. In every context, multiple actors are involved who all have their power, which may overlap between actors and the possibility of conflict. The system boundary includes the political, organisational and technological context and excludes the societal context, directly impacting the system as society has democratic rights. The judicial context is disregarded as they are independent. The decisions that are made are consequential and therefore resemble a decision-chain. The elaboration of the decision-chain discovers the unknown real-time outcome and complex processes that are part of this system. Information asymmetry adds to both and brings more uncertainty for the active actors. Additionally, the content changes independently of the design of the chain. The differentiation between characteristics of the whole system and those specifically for the decision-chain is significant when deepening the understanding of safety and eventually intervening in the system.

The central scientific gap denoted in this chapter is the lack of addressing the system components and characteristics of AI-supported decision-making in governmental organisations in the social domain and is filled through addressing the crucial system components in the defined system. Additionally, the first steps are taken to fill some other scientific gaps. These include the gaps towards the lack of direct feedback implementation in the socio-technical system of AI-supported decision-making towards the owner and higher organisational structure(s), lack of addressing and intervening on wickedness in an operational context, lack of addressing safety in a socio-technical system of AI-supported decision-making. The first steps are giving more insight into the system and addressing the main components.

The central empirical gap in this chapter is the lack of a proven understanding of the decision chain. This chapter presents the decision-chain in which it becomes clear what type of decisions are made at what stage. This includes the insight that the outcome is unknown in real-time and partly will be at any given future time. Thus, making the final decision and creating the outcome becomes crucial. Lastly, the differentiation in contexts helps to understand how decisions are made.

The main scientific gap in this chapter is addressing the system components, and characteristics of AI-supported decision-making in governmental organisations. The main empiric gap is the proven understanding of the decision-making system. The crucial system characteristics are denoted as:

- The interdependent contexts include the societal, political, organisational, and technological context, and finally, the judicial context on which the system is dependent.
- A multi-actor environment of only governmental actors within the system who have individual and overlapping power on how the decision-making system is structured.
- The system includes a decision-chain of which the final outcome influences (part of) society and includes two cycles, distinguishing this system from the socio-technical framework of Rasmussen (1997).
- The outcome is unknown in real-time and might only be partially known at any time, creating the inability to intervene in the outcome.
- Information-asymmetry among actors adds to uncertainties and conflict.
- The system objectives can differ over time, and content changes independently of the decision-making system, as balancing components are lacking.

The question arises how it is possible to mitigate unwanted system behaviour. How this system is kept safe is yet unknown. The system defined in this chapter serves as the base to create insight into the possible ways of regulation for safe system behaviour in the next chapter.

06 Safeguarding internal system safety

"How can internal system safety be understood appertaining to the system?"

Content

Takeaways

- Citizen's safety pinpoints both a scientific and empirical objective
- Due to the unknown outcome, the process ought to be controlled
- There are interdependencies between the system objectives and the system safety values
- As the objectives can change over time, the values might as well

6

Safeguarding citizen's safety

This chapter aims to define citizen's safety and integrate the notion with the framework designed in chapter 5. It is conducted by designing a conceptual system dynamics model that ought to capture the relations. This chapter legitimises the crucial system components and shows how the same concepts can be measured through differentiating contexts. Through the focus points addressed, there is added to the design cycle, depicted in figure 6.1.

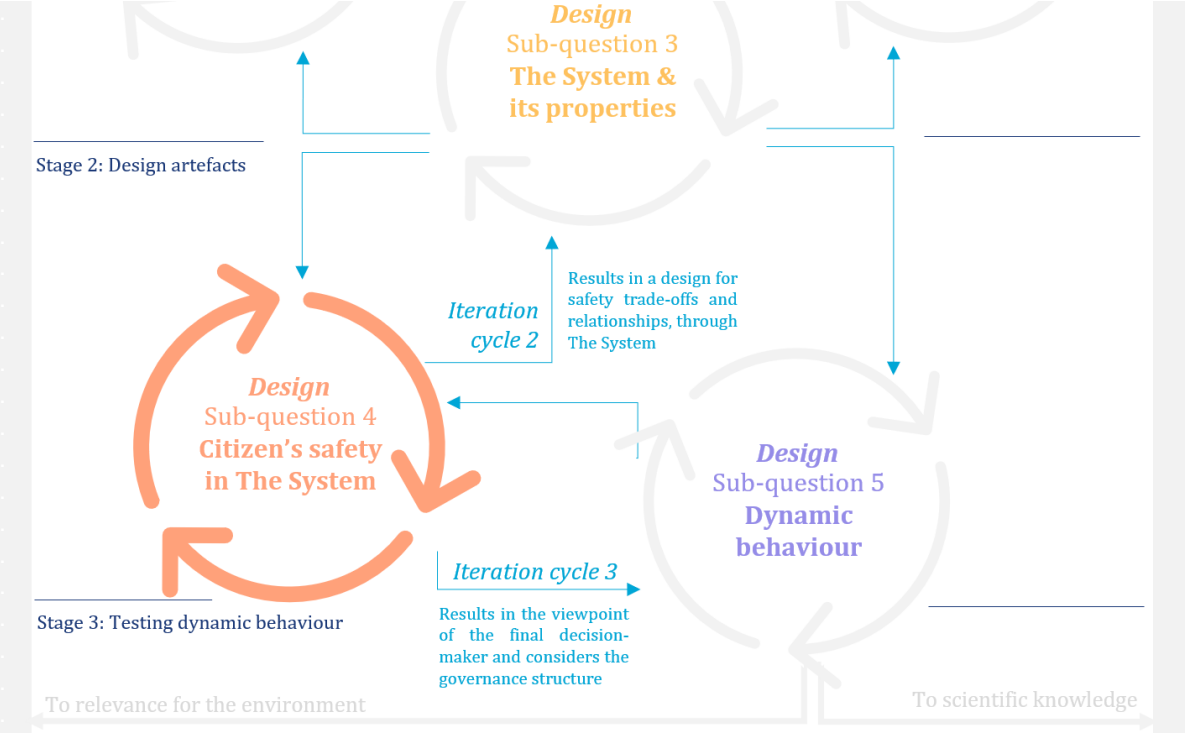


Figure 6.1: Design science cycles relevant for citizen's safety, inspired by Hevner (2014, p. 88)

Specifically, the sub-question "How can citizen's safety be understood appertaining to the system?" is answered in the following sections. Starting with defining the concept of citizen's safety, using the previous chapters as input to shape a meaningful definition in section **??**, followed by the definition of process-based citizen's safety in **??**. With these definitions, the object-value relationships become meaningful, shown in section 6.2, and the value dependencies, shown in section 6.3. Section 6.4 denotes the importance of dynamic behaviour when considering safety in an inherently dynamic system. Lastly, the conclusion is elaborated on in section 6.5.

6.1. From trustworthiness toward safety

Understanding the system gives insight into its different components, categorises them in different contexts, and comes with a crucial question. In a system where the problem is wicked and the outcome unknown, how can citizens be protected against harm?

Assessing interventions in the system is challenging because system components are interdependent, actors have their own and overlapping decision-making power and reflect specific behaviour, and the system includes feedback loops. It is impossible to measure the effect of The System as the how-question is unanswered. In a world of perfect information, it would be known whether the system outcome was correct, yet in reality, it possibly takes years if there is ever an exact outcome.

Interventions can be executed on different domains in the system, like changing the interdependencies by redirecting connections in the chain towards measurable goals with known outcomes or more actively involving society in the decision-making for a higher satisfaction level. To reach such interventions, one should always consider the wished-for and unwished-for effects.

Scientific & empirical explorations

The goal of trustworthy AI knows many aspects where politics, organisations, and technology come together. However, the notion of trustworthy AI does confuse, as the notion direct toward the algorithm itself and that implicates that lines of code can be interpreted as trustworthy or not. Besides consideration of the algorithm itself, the other contexts need to be taken into account as well, following the sociotechnical systems definition of Rasmussen (1997). Additionally, N. G. Leveson (2011) argues that sociotechnical systems need to be assessed as a whole to include its emergent properties, and therefore not be assessed as individual components. The EU sees the notion in a broader context as well, although their policy is structured around trustworthiness (High-Level Expert Group on AI, 2019). Additionally, using the notion of trust can be subjective and can have a different meaning for each actor, even when the term is broadly defined (Benk et al., 2022). Taking all of these arguments into account, the system safety definition of N. G. Leveson (2011) and the sociotechnical system definition of (Rasmussen, 1997) cannot be adopted without encountering new challenges. The biggest difference between their theories and The System defined in this research is the political dimension that adds uncertainty and potential of conflict to the absolute definition of safety, as the political dimensions defines the safety. As is the case in this research, laws and regulation become a dynamic property instead of a static one.

The book of N. G. Leveson (2011) brings an understanding of the theory of system safety in sociotechnical systems. Certain characteristics of this theory are paramount to compare to The System and its safety in this research. In systems safety's theory, safety refers to a system containing absence of accidents, where accidents are defined as loss. Additionally, many cases are about the safety of decision-making humans or objects, explaining workplace or aviation safety. In these types of systems, accidents are often directly identified by decision-makers in the system. For example, when a worker dies due to toxins in the air or a space shuttle burning due to an explosion. The origin of the accident may not be assessed through a system component analysis, as the system contains emergent properties. Therefore, the system as a whole requires consideration, even though it often includes complexities and uncertainties. Threats to the system safety can originate from a lack of vertical integration, creating a lack of feedback, because the system includes dynamic behaviour. What safety means is objective, as regulators, thus the overall norms, are often placed outside the system boundary, creating a common understanding of safety, i.e. not killing factory workers. Additionally, the defence against unsafe conditions erodes over time. Even though safety might not be a priority, or there may be other reasons why systems are not designed and operated safely, there is a high-level and objective understanding of what safety means. To apply the characteristics of system safety in sociotechnical systems in the context of this research, comparisons are required. To some extent The System can be assessed in light of the systems theory, however, some crucial differences remain for the technique presented by N. G. Leveson (2011).

Safety in The System

The goal of The System is to assess whether someone is rightfully or unrightfully receiving and has to be operated while avoiding harm to citizens. What it means to harm citizens can objectively be assessed when the harm is straightforward, and therefore the norms are equal among all actors. When

harm is avoided, citizens are safe. Citizen safety is paramount to the actors in The System, as protecting citizens against their own decisions is established in the Constitution (2018). Therefore, this research shifts from the notion of trustworthiness to safety.

Thus, the question of what it means to have safety in The System is posed. Safety in this research is not the safety for controllers, operators, or other decision-making in the system, but the safety of the citizens in the system with no particular role other than being the subject to the system. It is the role of government to protect citizens from their own unrightful decisions, as argued aforementioned. First, to judge whether the system is safe, the term safety requires definition.

If the citizens are not encountering unrightful decisions, they are safe. Looking into the benefits case, that assumption is invalidated. The government had the right to research the citizen, to primarily stop the benefits transaction, and to ask for all the details they wanted, because the law was explicitly harsh. Therefore, the assumption that as long as government abide law, the decision-making is safe, is not true. Different than in the original definition of sociotechnical systems by Rasmussen (1997), the regulatory bodies are part of the system. That creates an extra layer of complexity, as the laws that are formed suddenly become a variable instead of a constant. Two types of laws that can still be regarded as constants are the constitution, as it is above regular law and has to be accepted also after a change of government, and international treaties, including the European Union, as international law is above state law.

Another way of regarding a safe system is if no citizen is wrongfully categorised. In that case, the safety is dependent on the categorisation of citizens. This seems logical, however the answer to this is unknown in reality, as assessed in the previous chapter. Therefore, providing feedback on whether citizens are identified correctly or not, is impossible to give. It is, however, possible to steer the outcome in the most obvious ways, like checking whether someone actually receives a transaction from the government. These type of measures do not prevent harm to citizens as in theChildcare Benefits Case.

Defining safety includes a trade-off. On one hand, the wish is to avoid the harm in the system to citizens. One of the solutions can be to avoid controlling whether one is rightfully receiving, thus eliminating The System. In this way, service can be provided and no one would be subject to unrightful research, yet can receive when they need to. Harm cannot be done, hence safety is safeguarded. However, the effect of eliminating The System is that it is easier to commit unrightful transactions. When that happens, the system of providing service will flood and less money will be available for those who really need it, hence creating an unsafe situation as the benefits system either cannot be paid or the benefits allowance becomes too low to be beneficial to the citizens it regards. On the other hand, protecting the public money from fraudulent citizens is another definition prone to safety. Keeping the rightfully receiving citizens safe through detecting criminals seems logical, and for over ten years was deemed as being safe during the Childcare Benefits Case. Both ways of safeguarding safety can be argued for, yet both do no entail the complete avoidance of harm to citizens in the system.

Interestingly, this approach to safety shows wicked characteristics, as the solutions for safeguarding safety are not objectively true or false, but may be better alternatives than others and the solution again contains a new challenge.

The way forward

In the previous chapter, it is argued that the outcome of the system is unknown in real time. Due to this characteristic, it is impossible to implement real-time feedback that regards the outcome. Therefore, another way of organising feedback is required, which is the process of the system. The implemented feedback loops ought to balance The System.

In chapter 3 regarding the scientific exploration, governmental AI-supported decision-making is regarded in terms of challenges and in which many scholars argue for the safeguarding of privacy, transparency and prevention of biases. In the next chapter, 4, regarding the practical exploration, trustworthy AI, inter alia, is defined by privacy, transparency, and non-discrimination, whereupon the legal basis of these values in government is elaborated. In chapter 5, the unknown real-time outcome is supported by the wicked characteristic that there is not one utter solution nor binary expression of outcomes of the decision-making chain.

These findings altogether make for a conclusion that solutions in safeguarding the system cannot be sought in the real-time outcome yet can be sought in the process characteristics of the decision-chain.

The characteristics are of colossal influence and can be found quickly in the previously mentioned challenges. An ideal process can be defined as a continuous correct balance between privacy, transparency, and equality at all process stages of the decision-chain and considering the different contexts.

However, two values have yet been left out which are not necessarily crucial for the process but are for the overall working of the system. These vital values are efficacy and efficiency. Efficacy because the system must work correctly, even though the outcome is partially unknown. Efficiency, because (1) there is a certain capacity that is impossibly to do manually because (2) there is a limitation in means.

As privacy, transparency, and equality have a legal definition, the high-level understanding is already known. However, what it means to have those values when arguing in a more operational situation remains undefined, as the trade-offs are not explicitly made. Additionally, there are always exceptions to rules, and acts leave room for interpretation.

An illustrative example on the importance of a holistic view is the trade-off that can be made between the transparency of the algorithm and the bias of the stakeholders in the system, as more transparency of the algorithm creates more human bias. On the other hand, when the algorithm is transparent, a human can test its bias. However, transparency in the algorithm often reduces its efficacy, which directly influences citizens. This example shows why it is paramount to not only take transparency, equality, or efficacy into account, but to combine both social and technological aspects into one understanding.

6.2. Objective-value relationship

In the case of risk models for fraud detection, two values are directly involved: servitude and righteousness. Servitude, because the organisations are taking care of their citizens by transferring them the money they have the right to, which leads to righteousness. Control is needed to make sure the righteous people use the benefits. Righteousness plays a central role, yet servitude could be compromised. If it is known someone is a fraudster, do they have the right to the same servitude, or should they be judged strictly? Law, in general, gives space for interpretation to do precisely this: judge the nuances of what is ”fair”.

Dutch government (or its citizens) may not discriminate and must treat them equally. It might be wrongfully compromised when citizens are profiled and interpreted as discriminatory: treating one group differently from the other, which is insurmountable when wanting to detect fraud and selecting citizens to research based on their personal characteristics. However, it is legal to use profiling methods under certain circumstances, including fraud detection.

Values are mutually dependent and influence each other. It also concerns the previously lawfully embedded public values: anti-bias, transparency, and privacy. To make it relevant for this case, the trade-off between servicing citizens and fraud detection is incorporated, illustrated in 6.2.

The first relation describes that being more anti-bias leads to better servicing citizens, which is embedded in the constitution and a principle everyone needs to uphold, including government and its technology.

The second relation describes that being more transparent leads to better securing anti-bias. It is demonstrated by Felzmann et al. (2020), who argue that insight into the system leads to less discriminatory processes.

The third relation describes that being more anti-bias leads to less transparency, which is illustrated in the case of the social assistance benefit. In this case, data is pseudonymised, and the developer does not have access to personal details to prevent bias, thus making the data less transparent. One could argue that the developer could also not test for discriminatory outcomes in this way.

The fourth relation describes that enhancing privacy leads to more anti-bias. Protecting one's privacy results in the inability to use personal details for (unintended) discrimination, which is argued for in the GDPR.

The fifth relation describes that the more transparent government is, the better it can provide service to its citizens. A transparent government can explain their decisions and be open, which creates a more serviceable relationship (Felzmann et al., 2020).

The sixth relation describes that the better privacy is embedded, the harder it is to be transparent. This is because ensuring privacy can be an obstacle to giving full transparency. One example is the blacked-out dossiers in the child care benefit case (de Witt Wijnen, 2019).

The seventh relation describes that the more transparent government is, the harder it gets to detect

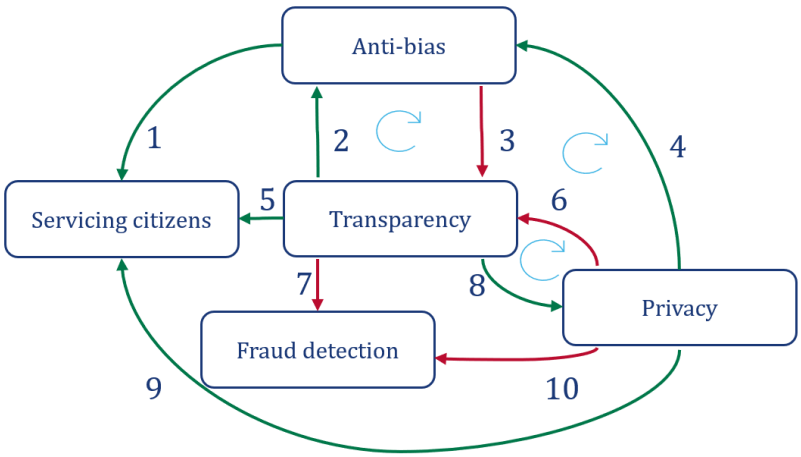
fraud because fraud is a crime of deception, which is easier when the one they try to defraud tells them how one is searching for them (Felzmann et al., 2020).

The eighth relation describes that more transparency leads to more privacy. Transparency inside the system leads to a better understanding of how to embed privacy, as only transparency processes can be monitored.

The ninth relation describes the more privacy; the better government provides service to their citizens, which is embedded through the GDPR, which becomes increasingly clear on how to embed privacy, in reality, (Felzmann et al., 2020).

The tenth relation states that the more privacy, the less fraud detection possible because a vast amount of data is needed to conduct data analysis and implement an AI. In the end, some of the data will be useless, which brings difficulty in compliance with the GDPR, which entails reasonable use.

Figure 6.2: Mutual dependent values in trade-off



The dependency in values shows the inter-dependencies not only among values but also among the different layers of the system. Suppose the organisational context cannot be transparent to the technological context. In that case, the chances of a good outcome become very small because the information is needed to design for privacy or against bias. It is evident that every layer has complexities and interacts with the surrounding layers. As research is lacking, assumptions must be made to differentiate between the different layers.

6.3. Value dependencies

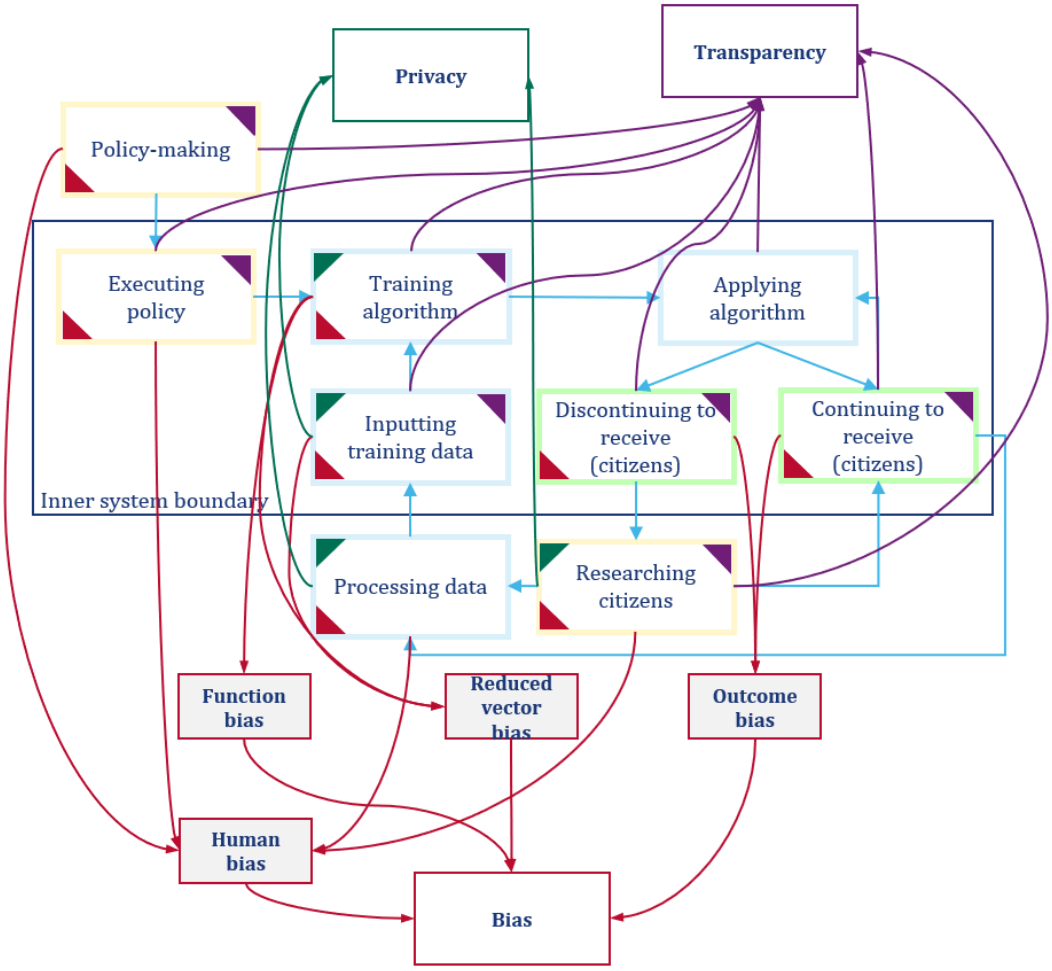
In the previous chapter, the system is defined, whereafter, in this chapter, the values are added to understand citizen's safety better. The values are relatively effortless and coupled within the inner system boundary. Interesting is that for most values, an equation can be used to calculate to what extent the value is present in the system, i.e. efficacy in section 6.3.5 and equality in section 6.3.2. However, not all values can be caught to that extent explicitly. The other values, i.e. privacy in section 6.3.3 and transparency in section 6.3.4, can increase through specific measures. How this comes together is illustrated in 6.3, and the four values are explained in the following subsections in context to this figure.

6.3.1. Considering figure 6.3

The colours previously used for the technological (blue) and organisational (yellow) context are used in this system subsequently. The demarcation is laid in the inner system boundary for this conceptual model, including the factors influencing or being influenced by the factors within the inner system boundary. As the inner system boundary signifies the boundary between the short and longer-term decisions and is far from the influence of society, this demarcation is chosen.

The triangles in the corners of the factors make the conceptual systems diagram more readable, as these are the factors that influence the outcome of privacy, transparency and bias. Training the algorithm, inputting training data, and researching citizens are the factors that influence all three values, which makes these the central nodes.

Figure 6.3: Conceptual systems diagram



For simplicity reasons, the figure does not contain any additional factors than those already present or linked to the values. However, the figure can be elaborated in multiple ways. The population is now left out, which can be used to calculate the efficacy of the final decision. Additionally, the money flows, or full-time equivalents, can be added to show the efficiency in terms of used means. For the latter, too little information is available to calculate the efficiency flows.

The arrows depict the causal relations that exist and therefore exhibit that all the factors that come in with an arrow are present for making the calculation. While the relations can be theoretically depicted, it is impossible to calculate them as specific values are unknown empirically.

The final decision-maker holds the power to put people back on receiving or not, influencing all of the values, thus a crucial player in the system. This person holds their power due to the human-on-the-loop governance structure to ensure a well-functioning system. However, it is unclear what kind of power this person holds empirically and theoretically. If this person is tasked with safeguarding citizen's safety, they need to be able to make changes to the processes or algorithm in order to ensure safeguarding. For this, feedback mechanisms must be in place to create a different value balance.

6.3.2. Equality

Equality in decision-making is important and incorporated into the constitution. Equality can be considered the opposite of bias and is not worth pursuing. Theoretically, bias can be calculated through the difference between the pre-existing and added biases. The formula is defined by Kleinberg et al. (2020), who also deconstruct the different forms of bias. The formula of the difference between added

and pre-existing bias ($D(f)$), resulting in $D(t \circ r)$:

$$D(t \circ r) = D(f) + (D(g) - D(r)) + (D(h \circ r) - D(g)) + (D(t \circ r) - D(h \circ r))$$

The terms all stand for a different part of the bias that is also defined in 6.3. The first term after the pre-existing bias, $(D(g) - D(r))$, refers to the difference in outcome measures when using a differentiated outcome measure compared to the actual outcome measure presenting the outcome bias in the illustrated conceptual system. The following term, $(D(h \circ r) - D(g))$, refers to the added bias as the vectors are reduced in the gathered data compared to the data in the population, which is illustrated as reduced vector bias. The last term, $(D(t \circ r) - D(h \circ r))$, refers to the bias added as the function for predicting is an estimated function from the actual function, illustrated as function bias. Currently, the calculated bias of the algorithm is unknown, as it is necessary to know the factors, in reality, to consider the differences rightly. However, it is currently possible to conduct sensitivity analysis and calculate the biases added from the training input until the model output, taking the training input as the population and the model input as the sample. Additionally, this formula shows that humans built the bias in algorithms, intentionally or not, which follows the last important factor to be considered in bias: human bias.

Kleinberg et al. (2020) argue that detecting human bias is even more difficult as human decision-making is neither transparent nor explicit compared to algorithm bias. Human bias has been an academically interesting topic for more than thirty years and is deemed complex (Jacob et al., 1986; Palframan et al., 2006; Peng et al., 2019). In the system presented in this research, human bias plays a central role in political and organisational decision-making, and to what extent remains unclear. In 6.3, the four bias types are connected to bias, as are the fraudsters and innocent in the population. The algorithmic bias resulting from the reduced factor, function, and outcome bias would also be known if they were known.

6.3.3. Privacy

Privacy is based on the used personal data of citizens in the system. However, there is no substantiated way to measure privacy other than when data links back to an identified person. Therefore, the measuring unit is unknown as well.

As the childcare and social assistance benefits cases did not show any constraints on including social security numbers and individual judgements of capabilities, it is assumed that concerning privacy, not many measures are taken. One measure is that the algorithm developers do not have access to social security numbers.

However, more technical measures exist to prevent linking the data to a person. Currently, the best performing measure is to add noise to the data to create differential privacy (Carvalho et al., 2022; Fletcher & Islam, 2019), which is a very technical approach to privacy and can make the training processes fully private. These measures are compelling and give an explicit definition of privacy, as the method consists of formulas. However, it is out of scope.

As a last remark, Carvalho et al. (2022) conclude that even with the method of differential privacy, the trade-off is made between the performance of the algorithm (efficacy) and the re-identification of individuals in the data (privacy).

6.3.4. Transparency

Transparency is illustrated in the middle of the training and model processes in figure 6.3. Transparency is about insight into the algorithm and has different forms. For example, source code can be publicised, i.e. open source, the workings of the algorithm can be publicised, i.e. partly in the algorithm registers Algoritmeregister (n.d.), the raw or metadata used for the algorithm can be publicised or a combination of these possibilities.

Transparency can be measured in overall contexts in certain cases (Hollyer et al., 2014); however, if it is unknown what might be happening, it cannot be known to what extent transparency is guaranteed.

Two remarks about transparency and its interdependencies with other values are that privacy is infringed with full transparency of personal information data sets. Additionally, when the decision rules of the algorithm are public, those who want to take advantage of the system are enabled, thus reducing efficacy. These arguments are supported by de Laat (2018) as well as by this research.

Transparency can be evaluated when more information is available on the childcare and social assistance benefits, e.g. in both cases, neither the code nor the decision rules are published. The

measures taken to ensure privacy, equality, and efficacy are neither published. Interestingly, a certain degree of transparency also exists in the other values. Currently, in both cases, this does not exceed stating that the values are guaranteed.

6.3.5. Efficacy

In the previous chapter, it is qualitatively concluded that the system's efficacy is complex to measure. The real numbers of people conducting fraud are unknown; it is unknown when one if one is researched, is innocent. In the lower part of 6.3, efficacy is illustrated and relates to the purple connections.

Figure 6.4: Formulas conceptual systems dynamic model Efficacy

Factor	Unit	Formula
Population	#[Citizens, SSN]	Given number
Fraudsters in Population	#[Citizens, SSN]	Population - Innocent
Innocent in Population	#[Citizens, SSN]	Population - Fraudsters
Model output	[-] (ordinal list)	Ranking[Citizens, SSN]
Research	#[Citizens, SSN]	First # of model output
Identified Fraudsters	#[Citizens, SSN]	Research – Id. Innocent
Identified Innocent	#[Citizens, SSN]	Research – Id. Fraudsters
Correctness Fraud	%	Count(IF(Identified Fraudsters[Citizen,SSN] IN Fraudsters[Citizen,SSN]))/Identified Fraudsters*100
Correctness Innocent	%	Count(IF(Identified Innocent[Citizen,SSN] IN Innocent[Citizen,SSN]))/Identified Innocent*100
Correctly Identified Fraudsters	#Citizens	IF(Identified Fraudsters[Citizen,SSN] IN Fraudsters[Citizen,SSN]): Correctly Id. Fraudsters += 1
Incorrectly Identified Fraudsters	#Citizens	IF(Identified Fraudsters[Citizen,SSN] NOT IN Fraudsters[Citizen,SSN]): Incorrectly Identified Fraudsters += 1
Correctly Identified Innocent	#Citizens	IF(Identified Innocent[Citizen,SSN] IN Innocent[Citizen,SSN]): Correctly Id. Innocent += 1
Incorrectly Identified Innocent	#Citizens	IF(Identified Innocent[Citizen,SSN] NOT IN Innocent[Citizen,SSN]): Incorrectly Identified Innocent += 1
Efficacy	%	(Correctly Identified Fraudsters + Correctly Identified Innocent)/ (Correctly Identified Fraudsters + Incorrectly Identified Fraudsters + Correctly Identified Innocent + Incorrectly IdentifiedInnocent) * 100

Efficacy is about how well the system performs. After the model output, the final decision-maker decides on the finite list of citizens to be researched, after which fraudsters and innocents are identified. As was previously argued, the innocent are only marked as innocent as they are not found guilty. To see whether this categorisation was proper, the citizens should be compared to the actual population lists. Unfortunately, this is theoretical. The exact number of fraudsters and innocent are unknown, yet the efficacy rate could be calculated if it was. The conceptual systems diagram distinguishes the population and the sample (citizens' data put into the technological context), from which the efficacy can be calculated.

The (in)correctly identified fraudsters or innocent are explicitly stated as it is possible to regard the two falsely categorised differently. For example, when striving for optimal fraud detection, false positives might be better than false negatives. Conversely, false positives might be judged as worse when striving to accommodate citizens optimally.

To calculate the factors, an explaining figure is given which describes the factors, units, and formulas accordingly in 6.4, specifically for efficacy. The arrows in the figure of **??** are defined through the formulas. The formulas partially validate the system and influence values, as the formula is (1) logical

and (2) the measurement units add up.

6.4. Dynamic objectives

In the childcare benefits case, it is observed that the Bulgarian fraud case emphasised detecting fraud in benefits. Because of this emphasis, the rules were interpreted strictly by all players, which resulted in strict law, an advising and judicial body that agrees, and others that want to follow the rules to prevent another such fraud case. This sphere of fraud detection as the highest goal shifted slowly when it was discovered that even the slightest mistakes of citizens had created their most considerable debts. When this came to light, the goal shifted toward servicing citizens, and the government was trying to recuperate by compensating said citizens. Even though, according to the previous harsh interpretation of the law, it is currently possible to conduct fraud with the recovery fund.

This example shows that the system is subject to dynamic changes and depends on the attitude towards the problem. This can be defined by the opposite of the former attitude, creating a sinus of the shift between goals. The trade-offs made by government employees are made differently in one context than the other. They are extra paramount when this involves influential actors, like the final decision-maker.

The values make it (im)possible to accomplish an objective, as shown in 6.2. Similar to the chicken or the egg, a causality dilemma can be posed. Do values create the possibility for objectives' existence, or do the objectives contribute to values' presence? Combining the latter interdependencies with the conceptual model in figure 6.3, and with section 6.3 results in a substantiated argument that all values are well connected. Including the three negative feedback loops between anti-bias and transparency, transparency and privacy, and all three values, theoretically, they should keep each other in check.

Therefore, even though the objectives change over time, the three values should be more or less balanced. Empirically this research has shown that when the objective is to detect fraud, the values of privacy and transparency will not be of the highest priority. Combining this information, it can be concluded that the system objectives, or at least detecting fraud, have a transcendent effect on the system values. It is also empirically shown that when focusing on detecting fraud, the priority was not on safeguarding equality. As there is no direct link between these, the effect of privacy and transparency on equality is assumed to be relatively large as well. These observations again show the importance of the final decision-maker, as it is their task to be the human-on-the-loop.

6.5. Conclusion

This section aims to conclude the results of this chapter and answer the fifth sub-question: "How can citizen's safety be understood appertaining the system?", answered by defining citizen's safety, mapping the inter-dependencies within the notion, and finally by elaborating on how the notion appertains the system.

Ideally, the system allows for continuous feedback on the status of citizen's safety. To continuously evaluate, the values need to be calculated continuously as well. Disregarding the unknowns present in the system is a complex task and urges more resources, perhaps making the system too inefficient. Efficiency is a value present in the background of this research, yet it is crucial as too much inefficiency suffocates the stakeholders. Disregarding the complications of the outcome, continuous feedback remains complex.

This chapter shows the central nodes in the system, containing the training algorithm, its input and the final decision through researching the citizens appointed by applying the algorithm. The first two factors are situated within the inner system boundary, referring to a faster schedule than outside. A good training algorithm and data are crucial to accommodate citizen's safety. When following logical reasoning argumentation, this conclusion complies.

Interestingly, one of the main factors, researching citizens, underlines the role of the implemented human-on-the-loop governance structure. One might question the differences regarding automated decision-making, disregarding this step. The question then is whether humans are safe to judge, i.e. human decision-making, or that humans, through technology, are safer to judge, i.e. algorithmic decision-making. One future scenario is that algorithms can be designed to judge values and behave accordingly. In this scenario, human decision-making might become controversial. Additionally, this scenario allows continuous calculation of values, making direct feedback into the system possible. It can be argued whether this approach is feasible.

This chapter answers the question, ”How can citizen’s safety be understood appertaining the system?”. Citizen’s safety is defined through the values of equality, privacy, and transparency. The values are mutually independent, as depicted in this chapter’s conceptual systems dynamics model. The conceptual model shows that the values are interconnected, so citizen’s safety aggregates these relations. The values crucial for citizen’s safety can be plotted on the system, as they derive from the decisions made at different stages. The dynamic behaviour is elaborated in the next chapter, and the conceptual systems dynamics model is validated through a semi-empirical experiment.

07 Testing dynamic system behaviour

"Is the dynamic system behaviour in line with the definition of the system components and internal safety?"

Content

Takeaways

- A game design complemented with a survey is designed
- Following the technological outcome is the opted choice when the final decision-maker lacks resources
- The conceptual system dynamics model is validated. Relationships among values exist

7

Dynamic system behaviour

The goal of this chapter is to show how dynamic behaviour is of influences safety and its values. This chapter is established through a serious game design and data analysis and validates the theoretical and empirical claims made in the previous chapter 6. This chapter ought to identify what dynamic behaviour means and can mean for the system defined in this research. Through the focus points addressed, there is added to all cycles, which is depicted in figure 7.1. Whereas the design cycle is added to the serious game design and data analysis, the relevance cycle generates semi-empirical results and the rigour cycle as the value-objective relationships are validated.

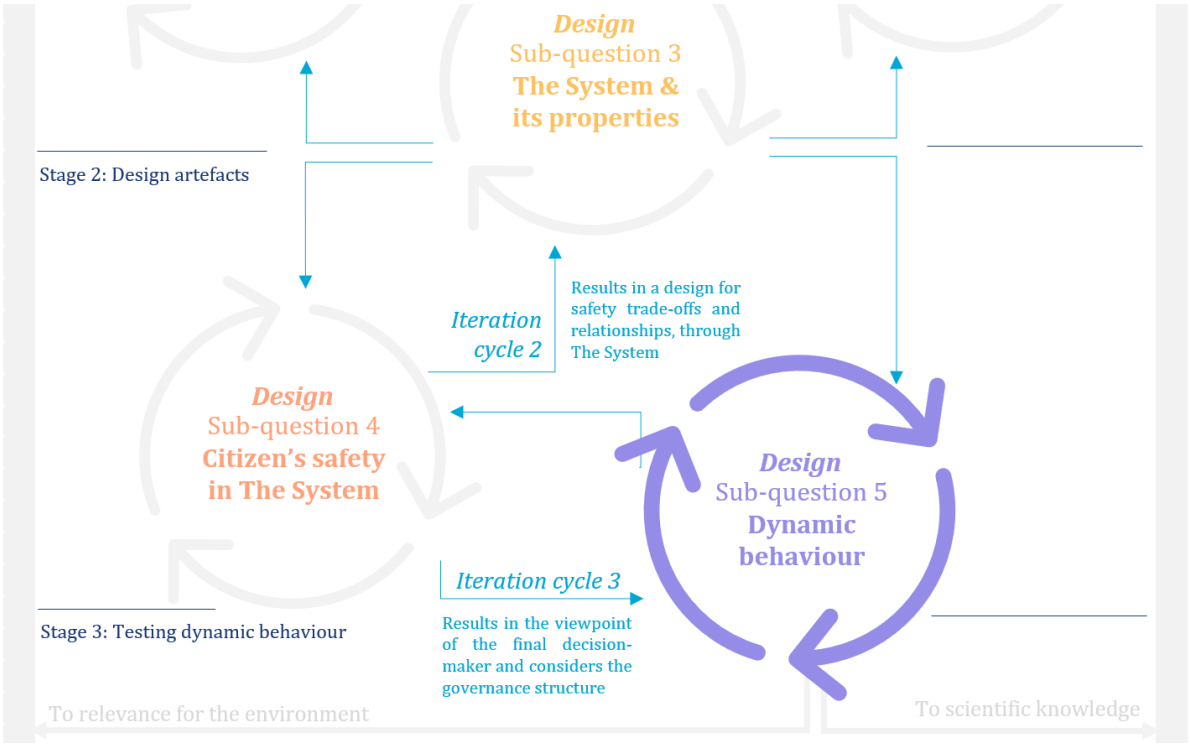


Figure 7.1: Design science cycles relevant for the dynamic behaviour, inspired by Hevner (2014, p. 88)

Specifically, the sub-question ”Is the dynamic system behaviour in line with the definition of the system components and safety?” is answered in the following sections, starting with the experiment set-up in section 7.1. The qualitative and quantitative results are presented in section 7.2. The conclusion is elaborated in section 7.3.

7.1. Experiment set-up

The goal of this section is to provide in-depth insights into the set-up of the experiment. Section 7.1.1 explains the demarcation of the experiment's set-up. Section 7.1.2 explain the choices made in the game design to give insight into how the game is construed. The following section, section 7.1.3 shows how the data is gathered and analysed.

7.1.1. Demarcation

The experiment is designed to function as a simple, serious game, including an additional survey to uncover the changing citizen's safety under differentiating system objectives. The scenarios presented in the game resemble the system, and the additional survey measures the differentiation between the two system objectives.

The external influences on the game are minimised through different aspects: The respondents are unaware of the tested objectives or the specifics of the research. The respondents were put in a scenario they could relate to but had probably not seen in their careers. Algorithms are not explicitly named, as this could enhance the prejudice in answering. Playing the game takes approximately 35 minutes.

The first scenario sketched in the experiment reflects the objective of detecting fraud: the so-called boss, the so-called new employee and the respondent. The boss represents the organisational boundaries, including the policy-making and executing stage of the system, as the boss decides on the policy ("catching the thieve") and the execution techniques (research by a colleague, final decision by respondent). The boss illustrates two more characteristics: the lack of time, coinciding with the need to make a decision depicted empirically, and the consequence, not being able to pay the salaries of his employees. The latter is an actual measure in the childcare benefits case, as the tax authorities had to collect the fraudulent money to pay their bills. The second persona, the new colleague, could not have stolen the money as they were not yet working for the said boss when the money was stolen. The colleague is not known by the respondent but did the research. Respectively, colleagues A, B, and C are regarded as high-risk, decreasing to a lower risk. The colleague is unknown as, in the system, the final decision-maker does not know in detail what is happening in the algorithm, which the colleague illustrates. Additionally, the colleagues suspected to have stolen the money stands for the model output or the high-risk citizens according to the score card in the system. It is emphasised that the respondent is to make the final decision solely and not with any (fictional) characters.

The second scenario knows the same characters: the boss, the new colleague, colleagues A, B, and C, and the respondent. This scenario differentiates from the first through the system objectives. In this case, the system objective is to be serviceable to citizens, corresponding with empiric research. The game completely turns around through speaking and the in-game objectives. Now it is the turn of the respondent to decide on a positive consequence: a promotion. The promotion is only in terms of money, to avoid interference of ideas like taking on more responsibility, or not fitting in a work-life balance. Receiving more money for the same work is generally viewed as positive.

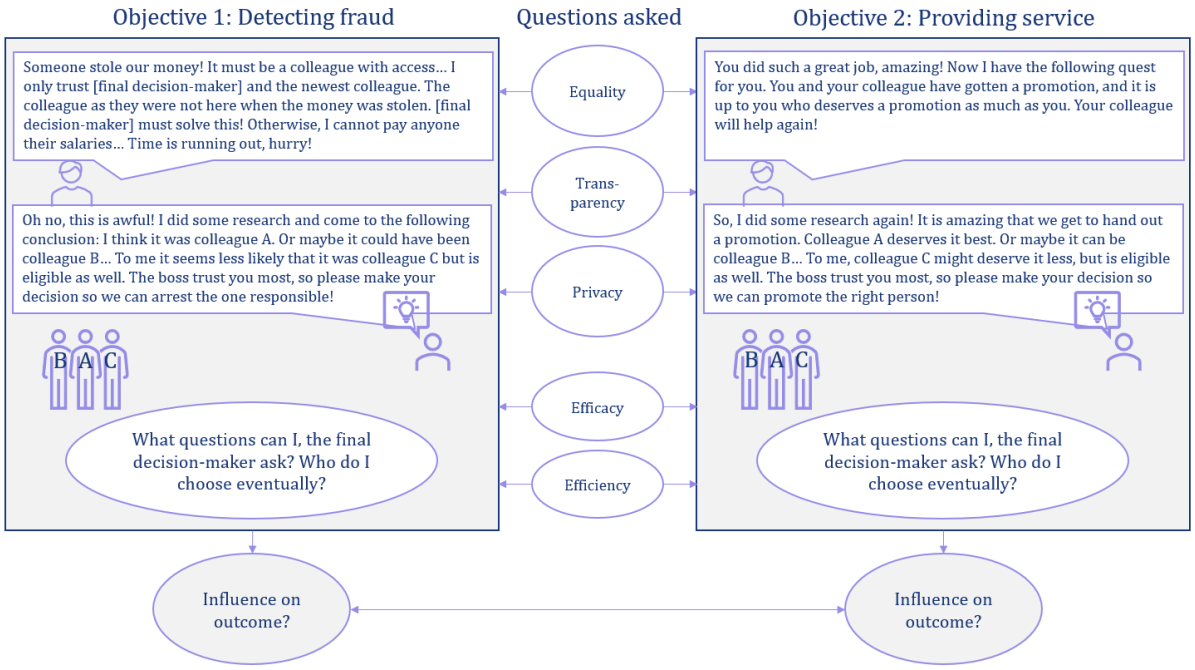
The respondent is urged to think out loud. Are there any questions the respondent would like to have answered? Why is one choice for a particular persona to be guilty? This part of the outcome is analysed quantitatively. The experiment set-up is shown in 7.2, which includes the two scenarios on the left, detecting fraud, and on the right, servicing citizens. At the bottom of the figure, the game output is depicted. The survey is depicted in the middle of the figure between the scenarios.

The survey consists of how one regards their decision-making, colleagues, and bosses for both scenarios. With this, the value-objective relations can be validated or nullified. The survey tests the correlation between values and objectives and the differentiation of importance for the organisational or technological context for the final decision-maker.

7.1.2. Choices in game design

Van Daalen et al. (2014) refer to the choices made in game design and are explicitly argued for in this section. To come to these choices is an iterative process, from designing the game to revisiting choices and coming to the conclusion that the choices made are not wholly reflected in the game. Two fully developed games have preceded the game presented in this research. The choices are defined by Van Daalen et al. (2014) as purpose; insight obtained, plot, players, roles, objective in-game/incentive, rules, representation of the physical system, and representation of the inter-actor environment. The

Figure 7.2: Set up of scenarios



following paragraphs discuss these choices to understand the final game design better. Furthermore, the game requires several practical choices: the time it takes to play the game, the number of people needed to bring together simultaneously to play the game, a simple explanation of the game and its rules, and non-complicated physical means in the game. These practical notions are learned through trial and error.

The objective of the game, the purpose, is to get insight into the decisions made by a final decision-maker when presented with not scarce information. It reflects the empiric situation, where the final decision-maker is the last stage in the decision-chain before impact on citizens occurs. As this governance structure is chosen primarily to reject automated decision-making, the final human decision-maker should have power in some way to control the outcome. Empirically, this decision-maker is not granted many means. None is known except for the check whether one receives, in this case, benefits, which cannot be regarded as a real influence on the outcome.

The insight obtained in the system is about the researcher obtaining the player's decision-making process as the player thinks out loud. The plot of the game is elaborated upon in the aforementioned section. Important to note is that the player always gets to hear that they did a great job in the previous scenario, which is why they get to do another scenario. This is done to stimulate the decision-making process. The players are experts in either policy analysis, technical models or both and have state-of-the-art knowledge about their fields as they are either working or graduates.

Further demographic characteristics are elaborated on in the next section. The role the player takes on is fictional yet close to reality. It is chosen to play a make-believe game to prevent the influence of the opinion after algorithm decision-making has been questioned in the public debate.

The objective of the game is to make the right final decision. The stakes are high regarding consequences for those (not) chosen. It stimulates the player to solute the problem, thus helping his colleagues, company and career. A comprehensive set of rules is already stated in the scenario sketches. The player can do two things: ask questions to the other personas in the game or conclude whom to pick. The answer to the questions does not contain more information than paraphrasing what is already conceived for the scenario sketch. Thus, interaction is minimised.

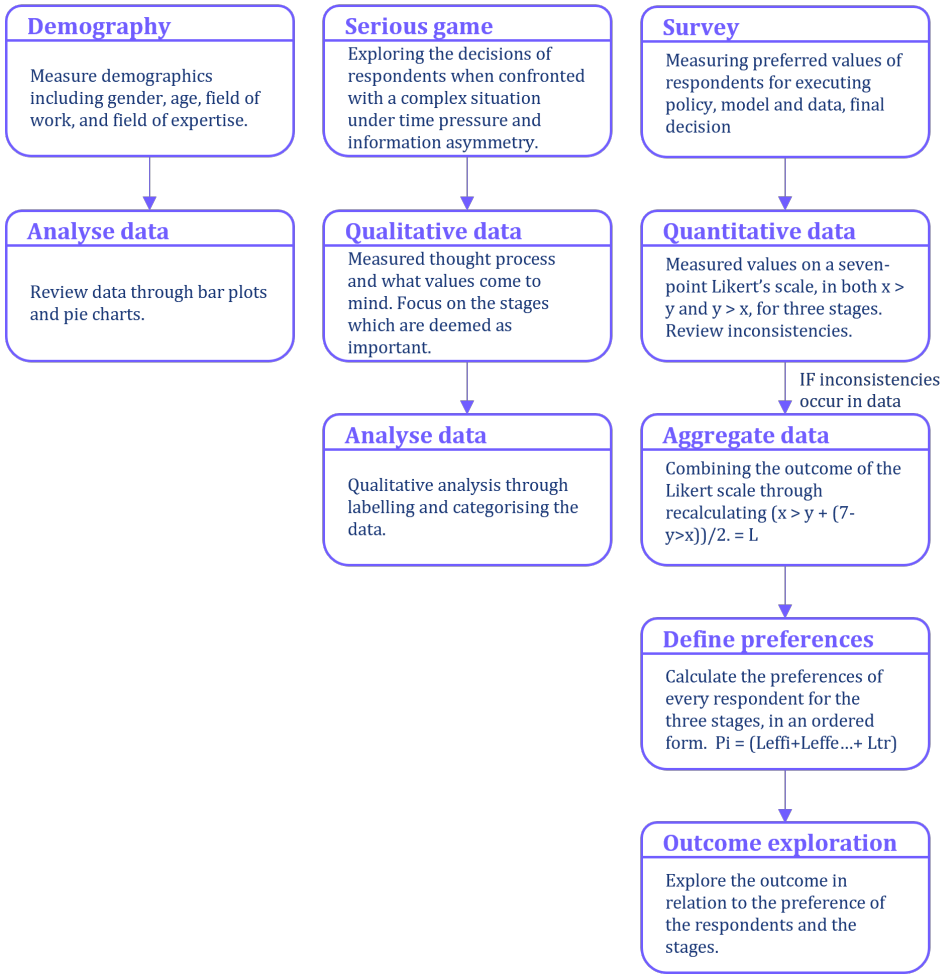
The game is played in a realistic environment in real-life with the researcher. However, the representation of the physical system is fictitious, as is the representation of the inter-actor environment. The latter is limited, as there is no real interaction, i.e. an underpinning decision tree.

As shown, the game depicted is relatively simple. This approach is chosen because the game is complementary to the research and not the primary method. Furthermore, in many serious games, the player learning a particular concept or aspect is the main objective for designing the said game; however, teaching the players something is not the objective of this research. The reason for using this method is that the game's results validate the two aforementioned chapters: the system and the conceptual systems diagram.

7.1.3. Data processing choices

This section aims to give insight into the process from data gathering to the outcome for the demography or respondents, the serious game and the survey. Notably, the three distinct data processes are designed to complement each other in answering the central sub-question of this research. Figure 7.3 illustrates the data processing process.

Figure 7.3: Data processing for demography, game, and survey



The demography is gathered through sex, age, the field of work and expertise. Important to note is that with sex, a diverse gaze is used for society, including the option for non-binary or inserting fluid options. The field of work aims to discover whether the respondent works in a public or private environment. Expertise is focused on whether one is in engineering/IT, policy analysis/decision-making, or different expertise. The demographic characteristics can be summarised in bar plots or pie charts.

The serious game results are analysed qualitatively. The game's results on the thought process before the respondents give the final answer is considered through labelling and categorising toward

a comprehensive overview of essential subjects in determining the final decision. Additionally, the outcome is noted here for both scenarios, with the reflection on the answers and whether the respondent would change their decision if they had the opportunity, with both the ideas that they answered correctly and incorrectly.

The survey is subject to the process with the most steps. It includes the quantitative data gathered from the survey. The survey asks about the respondent's preferences regarding values, differing from their preference for their own decisions toward their colleague and boss. The preferences are measured on a Likert scale, allowing for a composite interval scale (Joshi et al., 2015). The data is processed and summed for both x is more important than y and the other way around to prevent bias from human inconsistencies if inconsistencies are present in the data. The preferences per respondent are derived from seeing the relationship between the outcome and the preferences. This relation tells more about the relationship between values, stages and scenarios.

7.2. Experiment results

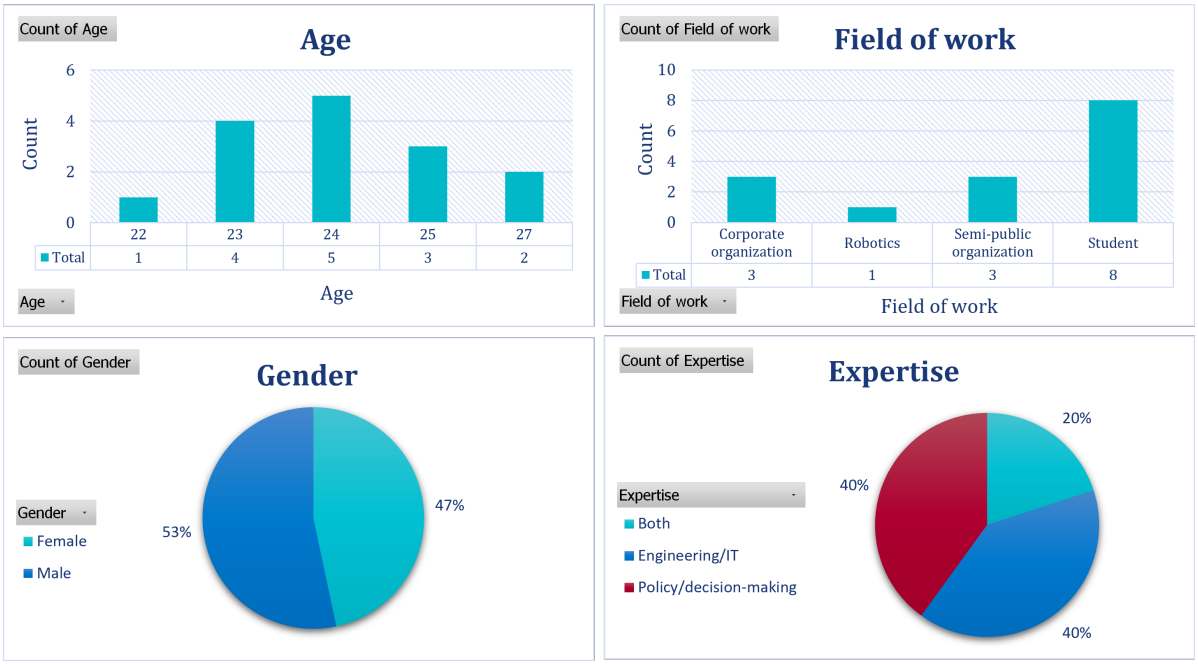
This section aims to present the results appropriately to serve the following section for concluding. For the illustration of the results, colourblindness and grey scales are considered. The results can be divided into three sections. First, the demographic characteristics of the respondents are visualised in section 7.2.1. Next, the qualitative results of the game are presented in section 7.2.2. Thereafter, the data analysis of the value-objective relationship is elaborated upon in section 7.2.3. This section is finalised with the preference ranking of values in section 7.2.4

7.2.1. Demography

Fifteen respondents answered four questions regarding their demographics, including their age, gender, field of work, and expertise, illustrated in 7.4.

In the graphs, it is noticed that the age is relatively young, under thirty. The gender between women and men is well-balanced. The field of work is added chiefly by students. However, both public and non-public organisations are represented. Expertise is perfectly balanced, as forty per cent represents respondents in engineering and policy or decision-making. The remaining twenty per cent is reserved for those with expertise in both.

Figure 7.4: Respondents' demography



7.2.2. Final decisions

The final decisions of both scenarios differ in most cases, even though the ranking stayed the same. In more than 60% of the cases, the respondent changes their choice in the following scenario. The final decisions are summed in the graph depicted in **??**. In the figure, it is seen that most respondents pick colleague A. It can be argued through the possibly perceived safety of the final decision-maker to stay with the recommended answer from the new colleague, in contrast to making their plan. Both pie charts depict that the difference between the scenarios is quite similar. However, in 7.6, it is shown that the respondents only stayed with their original choice five times, only when the original answer was A. Only in one case, A was not part of the answers.

Figure 7.5: Final outcome for both scenarios

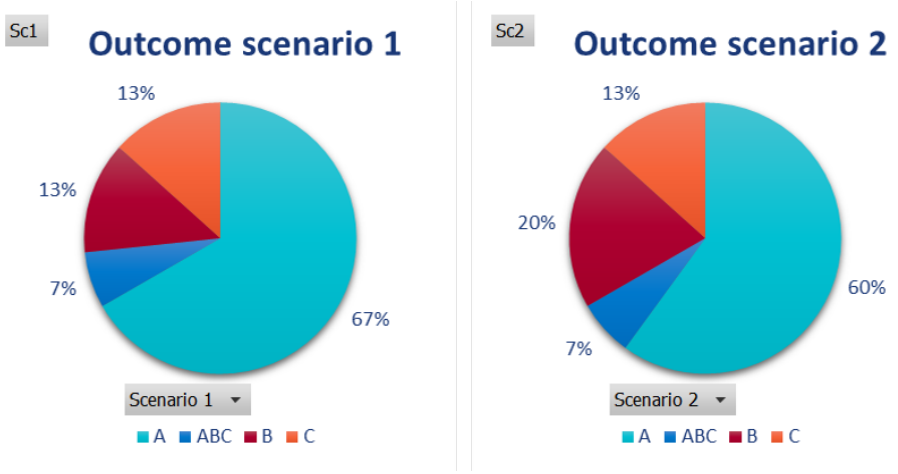
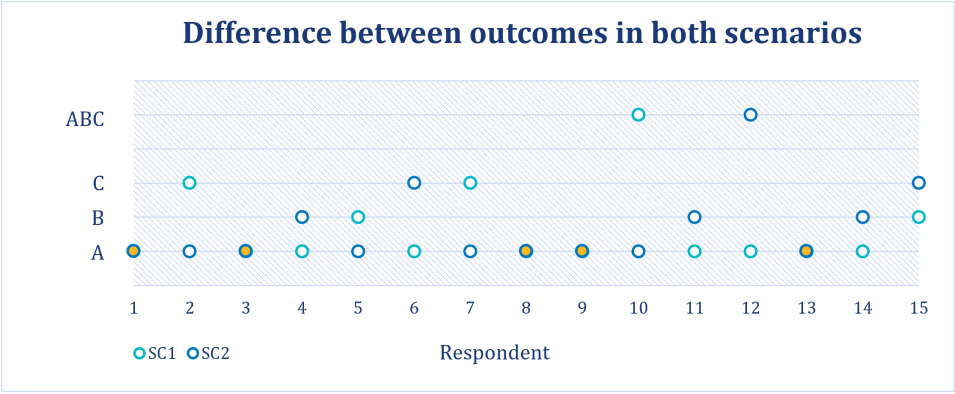


Figure 7.6: Final outcome for both scenarios compared



Whether one would make a different choice if confronted with the same situation is also measured. Once with the thought that one had given the correct answer (complete efficacy) and with the incorrect answer (no efficacy), which is depicted in 7.6. Interestingly, in twenty per cent of the cases, the respondent wanted to change their answer even when it was deemed correct for the first scenario, depicted in figure 7.7. This number increases to 33 per cent when the answer is thought to be false and is depicted in figure 7.8.

When told that the respondent was incorrect anyhow, more respondents wanted to change their answers. Unexpected is that respondents more often want to change their answer in the second scenario compared to the first. The difference is about forty per cent, almost half. It is unexpected as the consequences in the first scenario for the appointed colleague are severely negative (direct arrest), while in the second, it is not (increase in salary).

The insight into the reason for changing their decision is obtained qualitatively. It can be categorised into four main components: the respondent's trust in the data, their trust in their colleague, wanting to

Figure 7.7: Desire to change outcome when it is deemed correct

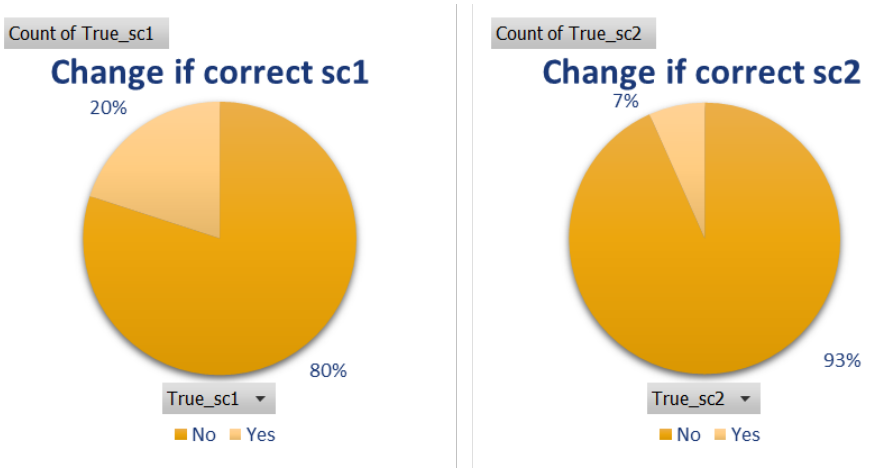
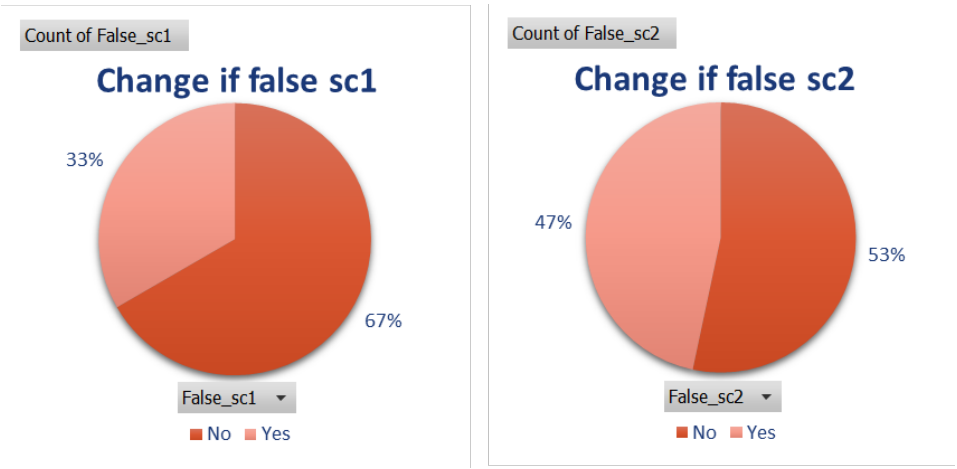


Figure 7.8: Desire to change outcome when it is deemed incorrect



make a more humane decision, and the desire to follow their plan. Interestingly, humane decision-making is present only if one thinks their answer was correct in scenario two and both false scenarios. It contradicts what is to be expected: emphasising humane decision-making when consequences are negative.

7.2.3. Value correlations

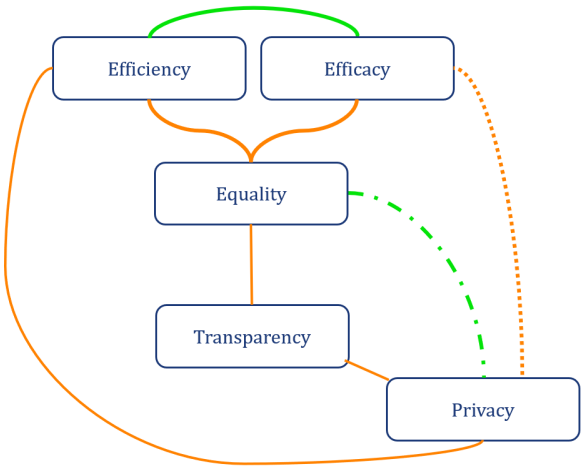
The goal of this section is to interpret the results for the value correlations. The results are obtained by extracting the inconsistencies. Thereafter, the data is aggregated from a Likert scale to a continuous one. For every new variable obtained, four are combined in one. For reviewing how this is done precisely, the script can be retrieved **here**. The correlations between values are depicted in figure 7.9 and serve as a validation for figure 6.2, and for which the data can be consulted in appendix A.

In figure 7.9 the relation between values is shown. Through correlation evaluation, it is not possible to see what variable influences what variable. Therefore, the arrows are replaced by lines. Bright green depicts the positive relationships between efficiency and efficacy and equality and privacy. Orange depicts negative relationships. Positive relationships mean that if variable A increases, variable B also increases. Negative relationships mean that if variable A increases, variable B decreases.

In the illustration, it is discovered that the negative relationships dominate, and the feedback loops occur between either one or three relations, meaning that the previously mentioned feedback loops are negative spirals. The thicker depicted relations (lines) relate to the significant relationships in both the general correlation matrix and the stage correlation matrices, depicted in appendix A. The dotted line (Efficacy - Privacy) shows that the relationship is seen in both matrices; however, when comparing the single-value matrix and the stage-dependent value matrix, it shows that the significance of this relation is only seen in the latter with $p<0.01$. It signifies that the relation does show but is not the strongest. It is contrary to the single-value matrix where in this relationship, $p<0.05$, meaning that the relationship is strong. It can be caused because the other variables in the latter do not control for the correlations. The stripe-dotted line (Equality - Privacy) is a relation only found in the stage-dependent value matrix.

The addition of efficiency and efficacy creates the option for validation. For example, if privacy increases, efficiency decreases, thus efficacy decreases, thus privacy increases. Another such loop is seen through privacy - efficiency - equality - transparency and privacy - efficacy - equality - transparency and privacy. Without considering these variables, only the loop equality - transparency - privacy - equality (and the other way around) could be checked out.

Figure 7.9: Value correlations



7.2.4. Ranking of values

This section depicts the importance of respondents when aggregated, shown in 7.10. The last row depicts the overall importance, the order being privacy, efficacy, equality, transparency and efficiency. Interestingly, efficacy is not the most paramount value in the overall importance, which may be caused by measuring the importance perceived by the final decision-maker. The final decision-maker may want their decisions private, so the boss cannot control them. It is a possible theory as, for the boss, privacy is only the fourth preference.

Referring to the tripartite division of principles by Prins et al. (2011), the driving principles (overall 2 and 5), the process-based principles (overall 4), and the underpinning principles (overall 1 and 3), the underpinning principles seem to weigh heavier than the other principles. Interestingly, for the final

decision-makers decisions (my decision) and the technological decisions (colleague's), the underpinning principles are deemed more important than the driving principles. At the same time, this switches when it is about policy execution (boss').

Figure 7.10: The values ranked in preference (high to low)

Ranking	**1**	**2**	**3**	**4**	**5**
My decision	Privacy	Equality	Transparency	Efficacy	Efficiency
Colleague's	Privacy	Equality	Efficacy	Transparency	Efficiency
Boss'	Efficacy	Efficiency	Transparency	Privacy	Equality
Overall	Privacy	Efficacy	Equality	Transparency	Efficiency

7.3. Conclusion

This chapter aims to answer the sub-question ”Is the dynamic system behaviour in line with the definition of the system components and safety?” to test the aforementioned synthesised knowledge. This question is answered through a serious game design, using qualitative and quantitative measures, the latter defined through the accompanied survey by the serious game. The answer to this question completes the cycle of design science, adding to the design science by designing a way for validating the preceding results and therefore giving back through the relevance and rigour cycle through a better understanding, respectively empirically and scientifically.

Both qualitative and quantitative measurements are used to detect, on the one hand, the way respondents come to a conclusion for a better explanation of the system and how the system would dynamically change semi-empirically. The choices made in game design are explicitly named to clarify the choice for the objective in the game and the objective of the game. Through a simple one-player game design, the game likewise meets the practical requirements. How the game’s data is analysed is depicted explicitly (7.3).

The system’s dynamic behaviour and safety are observed mainly through the outcome aggregations and the results regarding whether one wants to change their decision if the final answer was believed to be true or false. It would be expected that for both outcome scenarios, a difference can be seen in the outcome; however, it is minimal. Additionally, whether one answers A to the first or second scenario seems random. There is no distinction of order in the answers visible. What does stand out is that only one-third of the respondents is consistent with answering. Furthermore, when they are, the answer is A. And, the quantitative measurements through the survey allow for the possibility of validating the conceptual systems diagram constructed in the previous chapter, finding the relationships among values. Resulting in the possibility of assessing the presence of a relationship and, if true, whether it is positive or negative. In this, all value relationships are validated.

Thus, the answer to the sub-question ”Is the dynamic system behaviour in line with the definition of the system components and safety?” answers with a yes. However, that is not the only thing learned from this experiment.

One paramount conclusion regards the principle of human oversight. In chapter 3 the role of human oversight is introduced, including the governance structures HOTL, HITL, and HIC. From the empirical exploration, chapter 4, it becomes apparent that the HOTL principle is used in both policy and the benefits case. How the human final decision-maker is given decision-power through means is meagre. The human as the final decision-maker in the system creates an AI-supported system instead of an ADM, important as therefore the system is portrayed as safe by government.

When experimenting with the serious game, in which the final decision-maker can choose to follow the machine or not, they often do follow them. As the human as the final decision-maker is used for legitimisation and accountability for the decision that is made, this outcome is problematic for the regard of the benefits case. A liability remark is that the experiment is conducted explicitly with human decision-makers, to minimise the bias originating from taking into account that the alleged colleagues are defined through non-human decision-making.

In this experiment the conclusion of science on the malfunctions of human oversight is substantiated, from both a technical or operational perspective and from a social and political perspective. More respondents chose to go with the machine’s outcome when pressured and when the social climate was made to agree than in a scenario where the respondent was able to answer on their own believes, thoughts, and time.

The correlations between the values balance because the feedback loops are negative and check out. More privacy leads to less transparency. More equality leads to less transparency. The order of preferences for the values are interestingly observed when aggregating them to the scale of Prins et al. (2011). It is shown that in lower hierarchical decisions of the final decision-maker and the machine the underpinning principles or more preferred than the driving principles. On the contrary, for the organisational and tactical stage this shifts the other way around; driving principles are preferred over underpinning principles.

08 Discussion

Content

8

Discussion

This chapter elaborates on the discussion accompanying this research and is the first step in laying the groundwork for arguing the position of the research. The main question of this research is: ”How can citizen’s safety be safeguarded in governmental AI-supported decision-making?”. The research is exploratory, multidisciplinary and takes a systems perspective to the challenge. First, the research is reflected upon in section 8.1. Second and final, the recommendations for further research are depicted in section 8.2.

8.1. Reviewing results

Scientific and empiric exploration is executed to answer the main research question. The scientific exploration results in the definition of the required notions to give insight into the system. AI is defined broadly and includes a self-learning capacity. Therefore, predictive modelling lies in the scope. This technology’s foremost challenges include privacy intrusion, biases, and transparency. AI is a central notion in this research because it makes decision-making more complex for those executing the decision yet may prevent conflict, which relates to the wicked problem of the system. The decisions have a direct impact on citizens, demarcating the boundaries. Scientific consensus does not exist entirely on wickedness, yet the characteristics of high complexity, conflict, and uncertainty are agreed upon academically. This research uses wickedness to give insight into the system’s complexity, conflict, and uncertainty. The limitation is that the scientific exploration lacked the knowledge obtained empirically. Ideally, the literature validates understanding the system and its behaviour and maps pitfalls and boundaries. However, the scientific background is not matured to that stage, so other means of validation are sought.

Additionally, the main aim of this research is to obtain insight into how citizen’s safety can be safeguarded. Ideally, bias is prevented; thus, equality of citizens is guaranteed. Additionally, privacy is a right in the European Union, only granted with few exceptions. Lastly, transparency is based on the law, requiring the government to be as transparent as possible with its decision-making processes. Interestingly, the system at hand includes lawful exceptions on privacy, transparency, and even equality if interpreted in a certain way. Interpretation is that equality guarantees that two distinct groups are treated the same. However, in predictive modelling, the goal is to obtain differences between groups. Inequality is prohibited on the grounds of race, yet it is allowed on income. Two reasons for the importance of this are that (1) differentiating on the grounds of income or a combination of such variables can be a predictor for race, sex, sexual preference or other personal characteristics, and (2) that variables such as income can be seen as discriminatory as well, and is an ongoing discussion in society. Only these three values already make trade-offs around safety, as all these rights are there inter alia to protect citizens from government. The notion of citizen’s safety is designed to give a name for the safety this kind of trade-off entail.

Interestingly, the definition of citizen’s safety fits the definitions presented in the empirical exploration for trustworthy or responsible AI. However, the terminology does not fit what it truly entails. A technical software solution in itself is neither trustworthy nor responsible. What can be, is the system surrounding it, which includes the rules for execution, the reason it is used, who has access to it, and

the consequences of the output. Another differentiation is that trustworthy or responsible AI is also used for the safety regarding external factors, e.g. cyber attacks. Ultimately, citizen's safety researched in here covers the safety inside the system boundaries presented in this research, entailing a combination of values. Initially, system safety is a term used in engineering where technological appliances are used to do or make something, e.g. a factory. However, with the increasing importance of socio-technical systems, their presence in our everyday lives and the insufficient attention on safeguarding these systems, the need for safety socio-technical systems is born. Adding to this are the empirical examples shown in reality under the childcare benefits case. Even though the concept of system safety is not used as the political (conflict) dimension is exclude, the scientific term does lay the groundwork in exploring a new way of thinking.

Crucial in this research is the defined system, creating the foundation for further conducted research. The defined system consists of different contexts, stages and decisions synthesised from scientific and empirical exploration. The system depicts a decision-making chain compliant with Rasmussen's definition of socio-technical systems, explored in scientific exploration and used as the definition of a socio-technical system. Distinct in the system defined for governmental AI-supported decision-making is that this system also contains two feedback loops. Therefore, adding to the definition given in the scientific exploration.

Furthermore, the final decision is not made through the technical part of the chain (automated decision-making) but by the organisational one (AI-supported decision-making). It empirically resembles the current government structure in such systems and complies with privacy laws. It also means that the governance structure is inherent to the system. It is derived from the Human-on-the-Loop (HOTL) principle yet can be applied in cases of Human-in-the-Loop (HITL) and Human-in-Command (HIC). The system itself does not show any trade-offs. The only characteristics derived are wicked problems, multi-actor and multi-discipline, long or short term, and the effect on citizens. It is neither discussed nor illustrated what it means to make good decisions and add to minimise harm to citizens. Hence, citizen's safety follows.

The system is validated through interviews with professionals in the field regarding public decision-making or public algorithm use. The validation confirms the wicked problems and the challenges regarding AI. Additionally, it is clarified that this system is complex, with intricate and in-explicit trade-offs throughout every stage. In short, it is validated that the system deals with wicked problems and undefined safety concerns for citizens. An attached limitation of this validation is the hardship found surrounding knowledge gained with government employees working directly with the algorithms. The outcome directly impacts the citizens. The analysed documents are publicly available; however, finding the sought information required the exact correct search terms. One must know what one is looking for to obtain the wanted documents. Helpful are those who have analysed the documents and used the same notions or directly referenced them, whether in books, news articles or official reviews. The empirical information gained in this research is derived from official government documents and validated by professionals knowing government systems.

The system is used as a foundation for the citizen's safety. Both system and safety interdependencies are present. For the system, this means that the decision-making possibilities become more specific from every stage, starting with public opinion, politics, toward the organisation. Lock-ins are unavoidable, especially as developing technologies cost a lot of money. For citizen's safety, interdependencies occur among values. Even in a utopia, having a 100% of all values is impossible because they are contradictory. To create additional complexity, the requirements for safety may change over time. In general, the law is static. However, public opinion is not. Over time, the overall system objectives can change, as they did in the childcare benefits case over a decade, from complete fraud detection to complete service provision for citizens. Dynamic behaviour is paramount for citizen's safety, as all citizens have the same rights independent from the public will, yet it creates a challenge when decisions are time-dependent.

From the knowledge obtained until now, the values of equality, privacy, and transparency are connected, resulting in only one connection that is left out: equality toward privacy. All the other five connections are present. In this, the trade-offs and interdependencies between values are proved. When relating these to the system objectives of fraud detection or providing service to citizens, it becomes clear that providing service is emphasised with great equality and privacy, while with less privacy and

transparency, fraud detection is highlighted.

Adding the values to the different stages of the system gives insight into the central nodes regarding safety in the system. Central nodes train the algorithm and input the training data. Relating this to the earlier statement regarding trustworthy or responsible AI, the reasoning pertaining to the technology rather than the system becomes understandable. However, the training data and algorithm remain part of the system, shaped by the earlier decisions and having a certain degree of importance because of the subsequent decisions in the system.

Significant value relations are discovered between efficiency, efficacy, equality, privacy, and transparency. The relations obtained through the game and survey validate the conceptual systems diagram. Two characteristics are crucial to name in this discussion. First, the theory of values present in the decision-making in the system is depicted in the conceptual systems diagram. These relations are based on research previously conducted or synthesised from the knowledge obtained in this research. The survey attached to the game measures one's preferences from the perspective of the final decision-maker. What the respondent deems as important is measurement. It is different from measuring the values with more measurement methods and obtaining an overall idea and different from calculating the values. However, this is also not feasible as these methods contain many deep uncertainties. What is possible is the method presented in this research. It entails the assumption that, from a final decision-makers perspective, the first step in defining the means or interventions needed is to create a logical hypothesis for further research.

Additionally, this entails the consequence that measuring preference equals the decisions' values. It is possible as the assumption is that when one prefers a particular value over another, this is also depicted in one's decisions. For the quantitative research, the limitation must be that the total of respondents was fifteen. In literature, the consensus is thirty respondents when the population is homogeneous.

When asked whether the respondent wanted to change their decision knowing it was wrong, more than two-thirds answered no in the fraud case scenario, while slightly more than half of the respondents answered no in the service scenario. It is a peculiar outcome and might come across as unexpected. One possible explanation is that people dare to take responsibility if the stakes are not high, as the consequences are positive. However, they do not dare to diverge when someone else says it would be okay, even if that means the wrong person is arrested. This hypothesis is interesting to look into as it could explain why so many parents were wrongfully accused in the childcare benefits case.

8.2. Further research

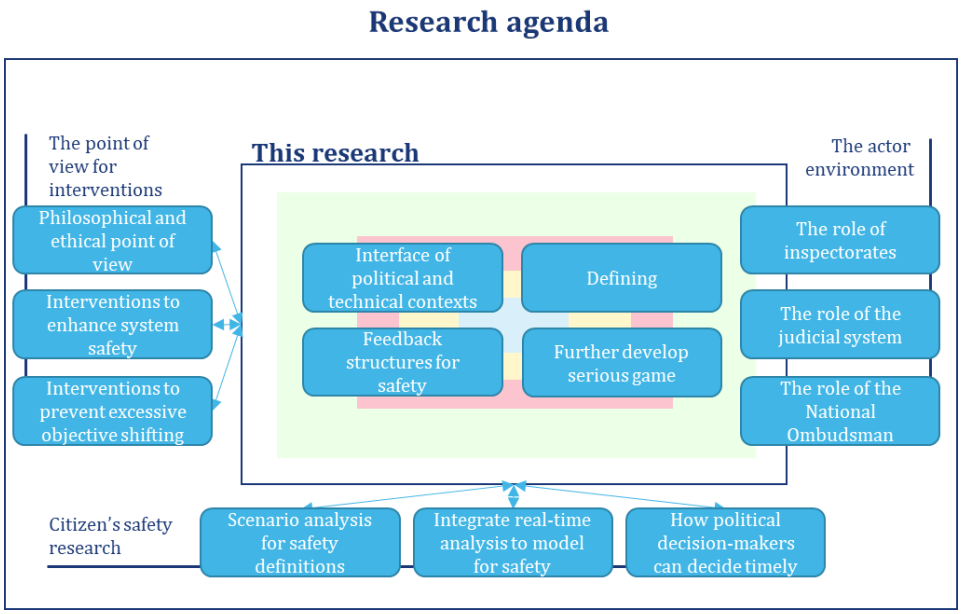
This section presents further research recommendations. The recommendations for further research are shaped according to the obtained knowledge and the notice of what is yet to be developed. Figure 8.1 shows the possible pillars for further research, after which three recommendations are highlighted.

The first suggestion for further research lies in the design of the system. The system, socio-technically defined by Rasmussen (1997) and for governmental AI-supported decision-making in chapter 5, does not define feedback loops upwards from the technological context toward the upper stages regarding policy-making or law-making. Ideally, feedback systems are identified and present. The suggestion for further research entails researching the possibilities of an upward chain. Currently, the decisions establishing the impact are shaped at the end of the decision-making process, of which the upper stages are left out. In one of two ways, this feedback loop is scoped in the Netherlands, left out of the scope for research.

The first argument regards the judicial system. Citizens can appeal the decision made about them in court, which rules in favour of citizens if they are rightful. The upside of this manner is that the court can protect citizens against the government. The downside of this manner is that appealing decisions can take years. It is time that the citizens whose benefits are stopped do not have, as most depend on benefits to pay the bill. Another downside of this manner is shown through the childcare benefits case, in which judges ruled against citizens. It is possible because laws are made in the system, and the court serves to interpret and rule on them. Hence, this manner of providing feedback into the system can serve citizens long term and the future citizens entering the system; however, it is not a solution for the current citizens.

The second regards the inspectorates. The goal of the inspectorates is to control and monitor the decisions made in the system. As of 2022, a new inspectorate is designed to control and monitor the

Figure 8.1: Research agenda



tax authorities (Official Gazette 2022-4749, 2022). The inspectorates control how the law is interpreted and how policy is executed. What they cannot do, is regulate to prevent harm to citizens if this is not unlawful. Like the judiciary system, the inspectorates are designed to monitor and control current laws, not shape new directions.

One authorised person to help citizens is the Ombudsman. Flaws in this system are the presumption that (1) citizens can access the Ombudsman and (2) the Ombudsman has the means to present the citizens' cases at the natural origin. Thus, the current way of feedback implementation lacks the protection of citizens when the law can sustain it. Understandably, protecting citizens and detecting criminals in the system, as presented in this research, is never a hundred per cent in reality. However, a way of direct feedback into the system serving to protect citizens while detecting criminals is not yet obtained. The first suggestion for further research is to obtain the solution spaces regarding how feedback may be designed in the system. The first step is to look into the existing feedback loops, determining the network of the people involved, which regards a more specific view than the presented actor analysis. Who knows who can be mapped through network analysis, yet it is an intensive measure regarding employees in such publicly sensitive systems. Therefore the research is to be conducted by a researcher with the connections to map this network. After mapping the network, including the means or power one has when presented with a citizen is discontinued by the system of its rightful receiving, it gives insight into the current direct feedback system and the central persons. Whether means and power have a working balance in the network can be reviewed.

The second suggestion for further research is about citizen's safety. Citizen's safety is currently defined through equality, privacy, and transparency. Research is conducted into the interdependencies between the values. It is shown how citizen's safety fits into the system and how values can be calculated in theory. These findings suggest further research in citizen's safety in three ways.

The first suggestion for further research into citizen's safety is the balance of when the system is safe. Citizen's safety is defined; however, what kind of balance between values is suited to the system objective is unclear. This research is conducted through an analysis of the will of the government stages and their goals. Additionally, the law is included to obtain information about the ongoing judiciary discussions on laws, their exceptions, and the philosophy of what should be. Experts in law, ethics, organisation and philosophy are required. To ensure the possible implementation, in reality, both an AI expert and field expert are included. The results entail when the socio-technical system can be regarded as internally safe and what circumstances explicitly add or subtract from the safety. The results of this research have a place in politics, the democratic core where laws are shaped.

The second suggestion for further research into citizen's safety is about reflecting on and assessing citizen's safety. This research follows the latter, as the results can be used to shape the scope of this second suggestion better. After the requirements for a safe system are clarified more in-depth, the notion can be reshaped to reflect better what it means to have an internally safe socio-technical system. Notably, the shaping of the notion is iterative. This research shapes the notion through the scientific literature, empirical exploration, and system definition. What is lacking is the refined iterations with reality to interpret reality adequately as a meaningful interpretation for science. The notion could be assessed quantitatively if the suggested research is conducted in iteration.

The third suggestion for further research into citizen's safety is about interventions. Intervening in the system and on citizen's safety is left out of the scope of this research. To come to meaningful conclusions on interventions, the notion of citizen's safety is assessed qualitatively and quantitatively. This way, scenarios can result from the intervention or a combination of interventions—one direction for defining the interventions lies in the governance structure. Governance structures are essential for the working of this system. However, the power of the final decision-maker, inherent to the structure through the type of decision-making, is not evaluated. The final decision-maker currently does not own explicit means, even though they are tasked with deciding on the impact on citizens. Interventions can look like equipping the final decision-maker with several means and analysing the impact on the citizen's safety.

The third and final suggestion for further research lies in the decision-making game, presented for the validation of value relationships. The game can be further developed to work with multiple players simultaneously, achieve different outcomes, and add a division of means to obtain specific information. In this way, the game is valid for learning about the decision-makers in the system and their impact on the system and obtaining data on the relationship between decision-making processes, citizen's safety, and the final outcome. In this way, the game can test as validation for the designed interventions in the system. It is suggested that by developing such an extensive game, serious gaming experts are involved, in addition to the researchers of the intervention development.

09 Conclusion

9

Conclusion

The goal of this chapter is to present a comprehensive conclusion, including key insights. The main research question is: ”How can citizen’s safety be safeguarded in governmental AI-supported decision-making?”. Through five sub-question, the answer to the main question is retracted, and beforehand the knowledge gaps are elaborated..

The specific scientific knowledge gaps addressed are the AI challenges from a holistic perspective, the collision or connection between political governance and operational safety, the operational wicked problem behaviour, and the feedback structures in governmental AI-supported decision-making. The specific empirical knowledge gaps addresses are the impact of Human-on-the-Loop (HOTL) governance structure, citizen’s safety in the system, values as a safety net, and the multi-actor environment. The gaps are addressed throughout the research and include a holistic perspective.

This research filled the gaps covering the holistic perspective. The main challenges for AI decision-making processes for citizens object to that process are the intrusion of privacy, equality, and transparency. This is in line with the notion of safe AI, for which the same values pursuit. Additionally, oversight structures depend on a human, which is empirically backed. These values can help in finding a safety balance in the system, as human oversight is not supported scientifically. This research substantiates that when errors are consequences from emergence, i.e. the decision-makers do not intentionally behave to create a false system outcome, the holistic perspective on the system helps identify the problem. As both the technical and social components are taken into account, the perspective adds not only to one pillar of the scientific world, but adds to the interconnection between disciplines. The interaction and connection between the different parts of the system is what characterises it and why it is a difficult system to assess. The challenges of AI interpreted over multiple disciplines, results in added value to the holistic view.

The investigation of collision or connection between political governance and operational safety is answered by the difference in approach and perspective. Where the theory of system safety adds to the operational stage and results in specific and measurable measures, the political governance is implicitly still debating on the definition of safety. As governance and other forms of law are born in the political arena, it becomes dynamic, including the notion of safety. Where in one coalition the ideas of safety include harsh enforcement to keep the general population safe, the other coalition strives for providing service to citizens. The collision between the operational and strategic level characterises this research.

Operational wicked problems are present in The System, characterised by the final decision directly impacting the environment of The System. The operational wickedness is observed and helps identifying the trade-off between detecting fraud and the providing service to citizens. It helps satisfying the lacking definite answer on the way interventions can be shaped and identifying the system and its characteristics. Although the impact of the decision is higher on the operation level than in policy, no definite changes to the problem are observed.

Feedback structures in governmental AI-supported decision-making are mainly present in the data used to train the algorithm. This poses a risk, as the values can be infringed. Other feed-back loops are not observes in The System. The Human-on-the-Loop (HOTL) governance structure is inherent

to the defined system, the loop it poses is not really a loop. The role of the final decision-maker is to control the algorithm's output, however, as seen in the cases, is given meagre means to uphold their task. Although the final decision-maker ensures that the AI-supported system is not an Automated Decision-Making (ADM), the loop they make is improbable.

As per the aforementioned paragraph, the HOTL structure's impact is sparse. The role of the final decision-maker in the system can be elaborated to include sufficient means to control the algorithm's decisions. However, the academic world doubts the capability of human oversight, which is substantiated by the experiment in this research. Therefore, government should search other ways of ensuring citizen's protection than with this governance structure.

Citizen's safety in the system is prone to implementation challenges. How the safety of citizens is implemented in the current system remains unclear. To obtain clearance the the political arena ought to define how they implement the citizen protection measures more explicitly, in line with the privacy laws and regulations. The organisational arena ought to consider how these values can be implemented. The last and final remark is on the current deemed unconsciousness about the implications, functionalities, and limitations of the technology used. Both the political and organisational arena need to be aware, respectively to strategical or tactical and to tactical or operational levels.

Values as a safety net may work when laws and regulations are elaborated by the European Union on equality, transparency, and other potential paramount values. As privacy is now protected internationally, the rules have strengthened for the allowances of ADM systems. As the solution may not be to put a human at the end of the decision-making process, more values require legal-based elaboration toward implementation.

As touched up in the aforementioned paragraphs, the multi-actor environment makes for a more complex system to intervene in. Concluding this gap is the way the boundaries of accountability function. Eventually, it is seen that shared responsibility is lacking responsibility. It is too easy to note that collaboration requires more priority, therefore, a clear division of accountable properties is argued for. When the rules for accountability are clear, one knows when one is responsible. This requires involvement of politics to make sure of logical decision-making boundaries. When wrong system outcomes exist nevertheless, the political arena should take their responsibility and step down.

The first sub-question is stated as follows: "What can be learned from previous research about safeguarding governmental AI-supported decision-making?". The answer to this question is obtained through a literature review resulting in scientific exploration. It is found that the notions crucial in this research are not precisely substantiated in definitions. Therefore the sought definitions are explicitly stated. AI is interpreted broadly, containing software with self-learning capabilities, including predictive modelling. Wicked problems contain problems with high complexity, conflict, and uncertainty, therefore resulting inter alia in a complex environment to intervene. Additionally a framework for socio-technical systems is presented, originally defined by Rasmussen (1997) and characteristics added derived from Dobbe (2022). This framework is the starting point for designing the system during sub-question three.

The second sub-question is stated as follows: "What can be learned about governmental AI-supported decision-making empirically?". The answer to this question is obtained through an empirical review, including official state documents. It is found that policy contains the will for trustworthy or responsible AI and that the Netherlands is leading in decision-making applications. Two cases are elaborated on and mainly show the gap between scientific and empirical exploration, where literature is not as mature as the systems implemented empirically.

The third sub-question is stated as follows: "What are the crucial system characteristics?". The answer to this question is obtained through synthesising the explorations and conducting a system and actor analysis. The system is defined through four contexts: societal, political, organisational and technical. The outcome flows from the technical context through the organisation to the societal context, the starting point of the analysis and system definition. The system is defined through decision-making stages that belong to a certain context and consists of a decision-making chain elaborated by two feedback loops. Therefore, it differs from the previously presented framework of socio-technical systems. Additionally, as aforementioned, two contexts are placed outside the system boundary, consisting of the judicial context that is not influenced by the system and the societal context. The crucial system characteristics obtained through the analysis are the unknown system's outcome, the technological output's uncertainty, information asymmetry among involved actors, wickedness in problem formula-

tion and therefore in the solution, and lastly, the influence of time resulting in the dynamic behaviour of the system and its objectives.

The fourth sub-question is stated as follows: ”How can citizen’s safety be understood appertaining to the system?”. The answer to this question is obtained through defining citizen’s safety and mapping the relations between citizen’s safety and citizen’s safety to the system and its objectives. Citizen’s safety is defined as serving the whole system. Citizen’s safety is an undefined term for socio-technical systems, therefore defined as the aggregated safety present in the system through the decisions made, resulting in a good balance between equality, privacy, and transparency. The definition of citizen’s safety takes the characteristics of governmental AI-supported decision-making into account through the defined values. A good balance remains undefined and can differentiate for different problems to the system is applied. Specifically, equality, privacy, and transparency are named to define citizen’s safety. Those values are a challenge in AI-supported decision-making and are embedded in law. The system characteristic time dependency is fundamental to citizen’s safety, as the system objectives can change over time. Therefore the pleasing balance of values for a safe system can also change. This dynamic behaviour of the system and citizen’s safety is conceptually defined by a conceptual systems dynamics model, resulting in an overview of interdependent relations. The value relations are tested through the next sub-question.

The fifth sub-question is stated as follows: ”Is the dynamic system behaviour in line with the definition of the system components and safety?”. The answer to the question of obtained through a simple, serious game design, including a one-player game and an additional survey. In the serious game, the respondent makes decisions being the final decision-maker in two scenarios, depicting the system objectives of criminal detection and providing service. The results show that most decisions align with what is suggested by the system depicted in the technological context. When asked whether the respondent wanted to change their decision knowing it was wrong, more than two-thirds answered no in the fraud case scenario, while slightly more than half of the respondents answered no in the service scenario. This peculiar outcome shows the importance of considering citizen’s safety. Another outcome of interest is validating that citizen’s safety in the system contexts is interdependent. The survey tested that the stages preceding the technological context, the technological context, and the stages succeeding this context are interdependent as they contain significant relations. In other words, the decisions made in these stages influence the citizen’s safety.

All of the sub-questions are crucial for answering the main research question. To provide an answer, the system and citizen’s safety required definitions. The system has been defined with the help of the first two sub-questions, coming together in the third sub-question. Those three sub-questions are needed to define the system and its characteristics; on the one hand, they result from the multi-disciplinary systems approach used in this research. Additionally, the wickedness - complexity, conflict, uncertainty - adds to the challenge of defining the system. Finally, the iterations of design cycle science provided the flexibility required to come to the definition and characteristics. The research has also helped shape the direction for defining citizen’s safety. The notion of citizen’s safety is constructed by mapping citizen’s safety on the system and the definition fitting all disciplines. In short, the answer to the main research question is that citizen’s safety can be safeguarded by and through the defined system.

Bibliography

2Doc. (2021). Alleen tegen de staat. https://www.npostart.nl/2doc/20-09-2021/BV_101407004

Aanwijzingen voor de regelgeving. (1992). BWBR0005730.

Agbozo, E., & Asamoah, B. K. (2019). Data-Driven E-Government: Exploring the Socio-Economic Ramifications. *JeDEM*, *1*(11), 2019. https://doi.org/https://doi.org/10.29379/jedem.v11i1.510

Algemene Rekenkamer. (2013). *Bezuinigingen op uitvoerings-organisaties* (tech. rep.). https://www.rekenkamer.nl/publicaties/rapporten/2013/01/24/bezuinigingen-op-uitvoeringsorganisaties

Algemene wet bestuursrecht. (1992). BWBR0005537.

Algoritmeregister. (n.d.). Overzicht van algoritmeregisters. https://www.algoritmeregister.nl/

Almada, M. (2019). Human intervention in automated decision-making: Toward the construction of contestable systems Machine Learning Regulation View project XAI in Tax Law View project Human intervention in automated decision-making: Toward the construction of contestable systems. https://doi.org/10.13140/RG.2.2.19766.55368/1

Amershi, S., Cakmak, M., Knox, W. B., & Kulesza, T. (2014). Power to the People: The Role of Humans in Interactive Machine Learning.

Amnesty International. (2021). Dutch childcare benefit scandal an urgent wake-up call to ban racist algorithms. https://www.amnesty.org/en/latest/news/2021/10/xenophobic-machines-dutch-child-benefit-scandal/

Autoriteit Persoonsgegevens. (2018). *De verwerking van de nationaliteit van aanvragers van kinderopvangtoeslag* (tech. rep.).

Autoriteit Persoonsgegevens. (2021). *Verwerkingen van persoonsgegevens in de Fraude Signalering Voorziening (FSV)* (tech. rep.).

Bakhshi, J., Ireland, V., & Gorod, A. (2016). Clarifying the project complexity construct: Past, present and future. *International Journal of Project Management*, *34*(7), 1199–1213. https://doi.org/10.1016/j.ijproman.2016.06.002

Bankes, S. (2010). *Robustness, Adaptivity, and Resiliency Analysis* (tech. rep.). www.aaai.org

Bannink, D., & Trommel, W. (2019). Intelligent modes of imperfect governance. *Policy and Society*, *38*(2), 198–217. https://doi.org/10.1080/14494035.2019.1572576

Baxter, G., & Sommerville, I. (2011). Socio-technical systems: From design methods to systems engineering. *Interacting with Computers*, *23*(1), 4–17. https://doi.org/10.1016/j.intcom.2010.07.003

Belastingdienst. (n.d.). Het systeem Fraude Signalering Voorziening (FSV). https://www.belastingdienst.nl/wps/wcm/connect/nl/contact/content/het-systeem-fraude-signalering-voorziening-fsv

Benefits Act. (2022). BWBR0004043. https://wetten.overheid.nl/BWBR0004043/2022-07-01

Benk, M., Tolmeijer, S., von Wangenheim, F., & Ferrario, A. (2022). The Value of Measuring Trust in AI - A Socio-Technical System Perspective. http://arxiv.org/abs/2204.13480

Blakeborough, L., & Giro Correia, S. (n.d.). *THE SCALE AND NATURE OF FRAUD: A REVIEW OF THE EVIDENCE 1 Collated by Laura Blakeborough and Sara Giro Correia* (tech. rep.).

Board of Directors Benefits. (2021). *Gegevensbeschermings effectbeoordeling GEB Doelgericht ingrijpen op een aangevraagde toesiag M1354 Het Risicoclassificatiemodel* (tech. rep.). file:///C:/Users/z.vantetterode/Downloads/Risicoclassificatiemodel_Toeslagen_deel_1.pdf

Canes-Wrone, B., Herron, M. C., & Shotts, K. W. (2001). *Leadership and Pandering: A Theory of Executive Policymaking* (tech. rep. No. 3).

Capano, G., & Woo, J. J. (2017). Resilience and robustness in policy design: a critical appraisal. *Policy Sciences*, *50*(3), 399–426. https://doi.org/10.1007/s11077-016-9273-x

Carvalho, T., Moniz, N., Faria, P., & Antunes, L. (2022). Towards a Data Privacy-Predictive Performance Trade-off. http://arxiv.org/abs/2201.05226

CBS. (n.d.). Huishoudens nu. https://www.cbs.nl/nl-nl/visualisaties/dashboard-bevolking/woonsituatie/huishoudens-nu

CBS. (2022a). 13,6 miljoen kiesgerechtigden bij gemeenteraadsverkiezingen. https://www.cbs.nl/nl-nl/nieuws/2022/09/13-6-miljoen-kiesgerechtigden-bij-gemeenteraadsverkiezingen

CBS. (2022b). Kenmerken van gedupeerde gezinnen toeslagenaffaire. https://www.cbs.nl/nl-nl/maatwerk/2022/26/kenmerken-van-gedupeerde-gezinnen-toeslagenaffaire

CBS. (2022c). Ruim 1 miljoen kinderen met kinderopvangtoeslag. https://www.cbs.nl/nl-nl/nieuws/2022/28/ruim-1-miljoen-kinderen-met-kinderopvangtoeslag

Cerquitelli, T., Quercia, D., & Pasquale, F. (2017). *Transparent Data Mining for Big and Small Data* (Vol. 32). Studies in Big Data. https://doi.org/10.1007/978-3-319-54024-5{_}1

Choi, Y., Gil-Garcia, J., Burke, G., Costello, J., Werthmuller, D., & Aranay, O. (2021). Choi_(2021)_Towards Data-Driven Decision-Making in Government. *IEEE Computer Society*, *2020-January*, 2183–2192.

Compatibility law 2016. (2016). BWBR0039429. https://wetten.overheid.nl/BWBR0039429/2018-01-01

Constitution. (2018). BWBR0001840. https://wetten.overheid.nl/BWBR0001840/2018-12-21

Council of State. (2013). *Advice Council of State Law fraud & fiscality 2013* (tech. rep.).

De Geus, A., Wijn, G., Ross-van Dorp, C., & Donner, J. (2005). Wet Kinderopvang. https://wetten.overheid.nl/BWBR0017017/2005-12-29

De Nederlandse Grondwet. (n.d.). Zetelverdeling Tweede Kamer. https://www.denederlandsegrondwet.nl/id/vh8lnhronvx6/zetelverdeling_tweede_kamer

De Stefano, V. (2019). 'Negotiating the Algorithm': Automation, Artificial Intelligence and Labour Protection. *Comparative Labor Law & Policy Journal*, *41*(1). https://doi.org/http://dx.doi.org/10.2139/ssrn.3178233

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

de Laat, P. B. (2018). Algorithmic Decision-Making Based on Machine Learning from Big Data: Can Transparency Restore Accountability? *Philosophy and Technology*, *31*(4), 525–541. https://doi.org/10.1007/s13347-017-0293-z

de Witt Wijnen, P. (2019). Opnieuw ophef in toeslagenaffaire: fiscus 'lakt' dossier zwart. https://www.nrc.nl/nieuws/2019/12/11/fiscus-lakt-van-alles-weg-maar-nog-niet-genoeg-a3983489

Dobbe, R. I. J. (2022). System Safety and Artificial Intelligence. http://arxiv.org/abs/2202.09292

Duan, Y., Edwards, J. S., & Dwivedi, Y. K. (2019). Artificial intelligence for decision making in the era of Big Data – evolution, challenges and research agenda. *International Journal of Information Management*, *48*, 63–71. https://doi.org/10.1016/j.ijinfomgt.2019.01.021

Dutch Government. (n.d.). Dataregistratie van de Nederlandse Overheid. https://data.overheid.nl/

Enserink, B., Hermans, L., Kwakkel, J., Thissen, W., Koppenjan, J., & Bots, P. (2010). *Policy Analysis of Multi-Actor Systems*. Lemma.

Ensign, D., Friedler, S. A., Neville, S., Scheidegger, C., Venkatasubramanian, S., & Wilson, C. (2018). Runaway Feedback Loops in Predictive Policing *. *Proceedings of Machine Learning Research*, *81*, 1–12. https://proceedings.mlr.press/v81/ensign18a.html.

European Union. (2018). General Data Protection Regulation.

Farrell, R., & Hooker, C. (2013). Design, science and wicked problems. *Design Studies*, *34*(6), 681–705. https://doi.org/10.1016/j.destud.2013.05.001

Felzmann, H., Fosch-Villaronga, E., Lutz, C., & Tamò-Larrieux, A. (2020). Towards Transparency by Design for Artificial Intelligence. *Science and Engineering Ethics*, *26*(6), 3333–3361. https://doi.org/10.1007/s11948-020-00276-4

Fletcher, S., & Islam, M. Z. (2019). Decision tree classification with differential privacy: A survey. https://doi.org/10.1145/3337064

Frederik, J. (2021a). Hoe de compensatieregeling van de toeslagenaffaire gierend uit de hand loopt. https://decorrespondent.nl/12987/hoe-de-compensatieregeling-van-de-toeslagenaffaire-gierend-uit-de-hand-loopt/781314790113-098def69

Frederik, J. (2021b). *Zo hadden we het niet bedoeld* (1st ed.). De Correspondent.

Gailmard, S. (2012). *Accountability and Principal-Agent Models* * (tech. rep.).

Galetsi, P., Katsaliaki, K., & Kumar, S. (2019). Values, challenges and future directions of big data analytics in healthcare: A systematic review. https://doi.org/10.1016/j.socscimed.2019.112533

General Childcare Benefit Act. (2022). BWBR0017017. https://wetten.overheid.nl/BWBR0017017/2022-08-01

General Data Protection Regulation (GDPR). (2016). 32016R0679. https://eur-lex.europa.eu/eli/reg/2016/679/oj

German Federal Government. (2020). Artificial Intelligence Strategy of the German Federal Government.
Government of Canada. (2021). Digital Nations Charter. https://www.canada.ca/en/government/system/digital-government/improving-digital-services/digital-nations-charter.html
Government of Canada. (2022). Algorithmic Impact Assessment tool. https://www.canada.ca/en/government/system/digital-government/digital-government-innovations/responsible-use-ai/algorithmic-impact-assessment.html
Government of the Netherlands. (n.d.). Prohibition of discrimination. https://www.government.nl/topics/discrimination/prohibition-of-discrimination
Government of the Netherlands. (2018). The Constitution of the Kingdom of the Netherlands 2018.
Grol, C. (2022). Transportondernemer vroeg voor tonnen aan kindertoeslag aan. https://fd.nl/samenleving/1443959/transportondernemer-vroeg-voor-tonnen-aan-kindertoeslag-aan
Head, B. W. (2019). Forty years of wicked problems literature: forging closer links to policy studies. *Policy and Society*, *38*(2), 180–197. https://doi.org/10.1080/14494035.2018.1488797
Henley, J. (2021a). Dutch government faces collapse over child benefits scandal. https://www.theguardian.com/world/2021/jan/14/dutch-government-faces-collapse-over-child-benefits-scandal
Henley, J. (2021b). Dutch government faces collapse over child benefits scandal. https://www.theguardian.com/world/2021/jan/14/dutch-government-faces-collapse-over-child-benefits-scandal
Hevner, A. (2014). *A Three Cycle View of Design Science Research* (tech. rep.). https://www.researchgate.net/publication/254804390
High-Level Expert Group on AI. (2019). *ETHICS GUIDELINES FOR TRUSTWORTHY AI* (tech. rep.). https://ec.europa.eu/digital-
Hoekstra, M., Chideock, C., & Van Veenstra, A. F. (2021). *Quick scan AI in de publieke dienstverlening II* (tech. rep.). https://zoek.officielebekendmakingen.nl/stcrt-2022-4749.html
Holligan, A. (2021). Dutch Rutte government resigns over child welfare fraud scandal. https://www.bbc.com/news/world-europe-55674146
Hollyer, J. R., Peter Rosendorff, B., Raymond Vreeland, J., Beck, N., Bonica, A., Broz, L., Davis, C., Gandhi, J., Gilligan, M., Little, A., Nagler, J., Naoi, M., Skorupski, W., & Stone, R. (2014). Measuring Transparency. *Political Analysis*, *22*, 413–434. https://doi.org/10.7910/DVN/24274
Holmberg, L. (2021). *Human In Command Machine Learning* (Doctoral dissertation). Malmö University, Department of Computer Science and Media Technology (DVMT). https://doi.org/10.24834/isbn.9789178771875
House of Representatives. (2009a). *Financieel jaarverslag van het Rijk 2008* (tech. rep.). https://zoek.officielebekendmakingen.nl/kst-31924-1.html
House of Representatives. (2009b). Wijziging van de Wet kinderopvang in verband met een herziening van het stelsel van gastouderopvang. https://zoek.officielebekendmakingen.nl/kst-31874-3.html
House of Representatives. (2013). *Bijlage Cijfers over het eerste kwartaal van 2013* (tech. rep.). https://zoek.officielebekendmakingen.nl/blg-231377.pdf
Inspectie Justitie en Veiligheid. (2022). *Hoe ging de Jeugdbescherming om met gezinnen gedupeerd door de Toeslagenaffaire?* (Tech. rep.).
Inspectie Overheidsinformatie en Erfgoed. (2021). *De informatiehuishouding van Toeslagen* (tech. rep.).
Jacob, V. S. J., Gaultney, L. D., & Salvendy, G. (1986). Strategies and biases in human decision-making and their implications for expert systems. *Behaviour and Information Technology*, *5*(2), 119–140. https://doi.org/10.1080/01449298608914505
Johansson, R. (2007). On Case Study Methodology. *Emerald Insight*, 48–54. https://www.emerald.com/insight/content/doi/10.1108/OHI-03-2007-B0006/full/html
Jones, A. J., Artikis, A., & Pitt, J. (2013). The design of intelligent socio-technical systems. *Artificial Intelligence Review*, *39*(1), 5–20. https://doi.org/10.1007/s10462-012-9387-2
Joshi, A., Kale, S., Chandel, S., & Pal, D. (2015). Likert Scale: Explored and Explained. *British Journal of Applied Science & Technology*, *7*(4), 396–403. https://doi.org/10.9734/bjast/2015/14975
Kamp, H. (2011). Brief van de minister van sociale zaken en werkgelegenheid. https://zoek.officielebekendmakingen.nl/kst-31322-116.html
Keys, P. (1990). *System Dynamics as a Systems-Based Problem-Solving Methodology* (tech. rep. No. 5).

Kinsner W, Wang Y, & Zhang D. (2010). System Complexity and Its Measures: How Complex is Complex. *Advances in cognitive informatics and cognitive computing* (pp. 265–295). Springer-Verlag.

Kleinberg, J., Ludwig, J., Mullainathan, S., & Sunstein, C. R. (2020). Algorithms as discrimination detectors. *PNAS*, *117*(48). https://doi.org/https://doi.org/10.1073/pnas.1912790117

Koens, L., & Vennekes, A. (2021). *The scope of AI research in the Netherlands* (tech. rep.). https://www.rathenau.nl/en/science-figures/research-artificial-intelligenc...

Koops, B. J. (2021). The concept of function creep. *Law, Innovation and Technology*, *13*(1), 29–56. https://doi.org/10.1080/17579961.2021.1898299

Koulu, R. (2020). Human control over automation : EU policy and AI ethics. *European journal of legal studies*, *12*(1), 9–46. https://doi.org/10.1145/3359301

Leendertse, J. (2022). Zwaarst gedupperden toeslagenaffaire nog niet geholpen, systeemoplossing is er nog niet. https://nos.nl/artikel/2423709-zwaarst-gedupeerden-toeslagenaffaire-nog-niet-geholpen-systeemoplossing-is-er-nog-niet

Lernende Systeme. (n.d.). Map on AI. https://www.plattform-lernende-systeme.de/map-on-ai-map.html

Leveson, N. G. (2011). *Engineering a Safer World* (tech. rep.).

Leveson, N., Samost, A., Dekker, S., Finkelstein, S., & Raman, J. (2020). *A Systems Approach to Analyzing and Preventing Hospital Adverse Events* (tech. rep. No. 2). www.journalpatientsafety.com

Leveson, N. G. (2017). Rasmussen's legacy: A paradigm change in engineering for safety. *Applied Ergonomics*, *59*, 581–591. https://doi.org/10.1016/j.apergo.2016.01.015

Levy, K., Chasalow, K., & Riley, S. (2021). Algorithms and Decision-Making in the Public Sector. *Annual Review of Law and Social Science*, *17*, 309–334. https://doi.org/10.1146/annurev-lawsocsci-041221-023808

Li, N., Adepu, S., Kang, E., & Garlan, D. (2020). Explanations for human-on-the-loop: A probabilistic model checking approach. *Proceedings - 2020 IEEE/ACM 15th International Symposium on Software Engineering for Adaptive and Self-Managing Systems, SEAMS 2020*, 181–187. https://doi.org/10.1145/3387939.3391592

Liem, C., & Nasrullah, I. (2022). The Child Benefit Scandal Through Computer Scientists' Eyes.

Liu, H. W., Lin, C. F., & Chen, Y. J. (2019). Beyond state v loomis: Artificial intelligence, government algorithmization and accountability. *International Journal of Law and Information Technology*, *27*(2), 122–141. https://doi.org/10.1093/ijlit/eaz001

Lönngren, J., & van Poeck, K. (2021). Wicked problems: a mapping review of the literature. *International Journal of Sustainable Development and World Ecology*, *28*(6), 481–502. https://doi.org/10.1080/13504509.2020.1859415

Lum, K., & Isaac, W. (2016). To predict and serve? *Significance*, *13*(5), 14–19. https://doi.org/https://doi.org/10.1111/j.1740-9713.2016.00960.x

Makridakis, S. (2017). The forthcoming Artificial Intelligence (AI) revolution: Its impact on society and firms. https://doi.org/10.1016/j.futures.2017.03.006

Marda, V. (2018). Artificial intelligence policy in India: A framework for engaging the limits of data-driven decision-making. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *376*(2133). https://doi.org/10.1098/rsta.2018.0087

Ministerie van Financiën. (2021). *Bijlage 1. Indicatoren gebruikt bj het risicoclassificatiemodel voor KOT door de jaren heen* (tech. rep.).

Ministry of Internal Affairs, & Ministry of Justice and Safety. (2022). Wet open overheid. https://wetten.overheid.nl/BWBR0045754/2022-08-01/0/afdrukken

Ministry of Internal Affairs and Communication. (2019). *AI Utilization Guidelines* (tech. rep.).

Ministry of Social Affairs and Employment. (2021). *Introductie dossier Ministerie van SZW* (tech. rep.).

Municipality of Rotterdam. (n.d.). Algoritmeregister. https://www.rotterdam.nl/bestuur-organisatie/algoritmeregister/

Munir, S., Stankovic, J. A., Liang, C.-J. M., & Lin, S. (2013). *Cyber Physical System Challenges for Human-in-the-Loop Control* (tech. rep.).

Nahavandi, S. (2017). Trusted Autonomy Between Humans and Robots: Toward Human-on-the-Loop in Robotics and Autonomous Systems. *IEEE Systems, Man, and Cybernetics Magazine*, *3*(1), 10–17. https://doi.org/10.1109/msmc.2016.2623867

National Ombudsman. (2010). Terugvordering kinderopvangtoeslag 'De Appelbloesem'. https://www.nationaleombudsman.nl/nieuws/2010/terugvordering-kinderopvangtoeslag-de-appelbloesem
Nationale Ombudsman. (n.d.). De Nationale Ombudsman. https://www.nationaleombudsman.nl/de-nationale-ombudsman
OECD. (2019). *The Path to Becoming a Data-Driven Public Sector*. OECD Digital Government Studies. https://doi.org/10.1787/059814a7-en
Official Gazette 2022-4749. (2022). Regeling van de Minister van Financiën van 9 februari 2022 (nr. 2022-39462), houdende wijziging van het Organisatiebesluit Ministerie van Financiën 2020 en het Mandaatbesluit Ministerie van Financiën 2020 in verband met diverse organisatorische wijzigingen.
Organisatiebesluit Ministerie van Financën 2020. (2022). BWBR0043027. https://wetten.overheid.nl/BWBR0043027/2022-07-20
Overheid.nl. (n.d.). Wat doet de overheid? https://www.overheid.nl/zo-werkt-de-overheid
Palframan, W. J., Meehl, J. B., Jaspersen, S. L., Winey, M., & Murray, A. W. (2006). Anaphase inactivation of the spindle checkpoint. *Science*, *313*(5787), 680–684. https://doi.org/10.1126/science.1127205
Paola, L. (2021). Bias does not equal bias: A socio-technical typology of bias in data-based algorithmic systems. *Internet Policy Review*, *10*(4). https://doi.org/10.14763/2021.4.1598
Parlement. (n.d.). Kabinetsformatie sinds 1945. https://www.parlement.com/id/vh8lnhrsr2z2/kabinetsformaties_sinds_1945
Peng, A., Nushi, B., Kıcıman, E., Inkpen, K., Suri, S., & Kamar, E. (2019). What You See Is What You Get? The Impact of Representation Criteria on Human Bias in Hiring. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, 125–134. www.aaai.org
Prins, C. (, Broeders, D., Griffioen, H., Keizer, A., & Keymolen, E. (2011). *iGovernment* (tech. rep.). The Netherlands scientific council for government policy. Amsterdam University Press. https://english.wrr.nl/publications/reports/2011/03/15/igovernment
PWC. (2022). *Onderzoek effecten FSV Voorbeeld casussen* (tech. rep.).
Rahwan, I. (2018). Society-in-the-loop: programming the algorithmic social contract. *Ethics and Information Technology*, *20*(1), 5–14. https://doi.org/10.1007/s10676-017-9430-8
Rasmussen, J. (1997). Risk management in a dynamic society: a modelling problem. *Elsevier*, *27*(2/3), 183–213.
Rechtbank Utrecht. (2010). ECLI:NL:RBUTR:2010:BL0008. https://uitspraken.rechtspraak.nl/inziendocument?id=ECLI:NL:RBUTR:2010:BL0008
Regulation implementation Tax Authority. (2022). BWBR0014506. https://wetten.overheid.nl/BWBR0014506/2022-01-01
Rijksoverheid. (n.d.-a). Taken van de rijksoverheid. https://www.rijksoverheid.nl/onderwerpen/rijksoverheid/taken-van-de-rijksoverheid
Rijksoverheid. (n.d.-b). Wanneer heb ik recht op kinderopvangtoeslag?
Rijksoverheid. (n.d.-c). Wat zijn de taken en bevoegdheden van de belastingdienst? https://www.rijksoverheid.nl/onderwerpen/inkomstenbelasting/vraag-en-antwoord/wat-zijn-de-taken-en-bevoegdheden-van-de-belastingdienst
Rijksoverheid. (n.d.-d). Wie mag er stemmen bij de gemeenteraadsverkiezingen. https://www.rijksoverheid.nl/onderwerpen/verkiezingen/vraag-en-antwoord/wie-mag-stemmen-bij-gemeenteraadsverkiezingen
Rijksoverheid. (2020). Belastingdienst Toeslagen start campagne kinderopvangtoeslag: Verandert er iets? Geef het meteen door. https://www.rijksoverheid.nl/actueel/nieuws/2020/09/11/belastingdienst-toeslagen-start-campagne-kinderopvangtoeslag
Rittel, H. W. J., & Webber, M. M. (1973). *Dilemmas in a General Theory of Planning* (tech. rep. No. 2). https://doi.org/https://doi.org/10.1007/BF01405730
Roobeek, R., Frater, J., & Kennedy, N. (2021). Dutch government resigns over child welfare fraud scandal. https://edition.cnn.com/2021/01/15/europe/netherlands-government-resigns-scandal-intl/index.html
Rotterdam Court of Audit. (2021). *Gekleurde technologie* (tech. rep.). https://rekenkamer.rotterdam.nl/wp-content/uploads/2020/11/R.P.20.06-gekleurde-technologie.pdf
Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., & Fei-Fei, L. (2015). ImageNet Large Scale Visual Recognition

Challenge. *International Journal of Computer Vision*, *115*(3), 211–252. https://doi.org/10.1007/s11263-015-0816-y

Saalfeld, T. (2000). *Members of parliament and governments in Western Europe: Agency relations and problems of oversight* (tech. rep.).

Scopus. (2022a). Analyze search results. https://www.scopus.com/term/analyzer.uri?sid=dc2ec386c9a95d02a02a618815faf48c&origin=resultslist&src=s&s=TITLE-ABS-KEY%28wicked*+W%2f5+problem*%29&sort=plf-f&sdt=a&sot=a&sl=35&count=1651&analyzeResults=Analyze+results&cluster=scopubstage%2c%22final%22%2ct%2bscosubtype%2c%22ar%22%2ct%2bscolang%2c%22English%22%2ct&txGid=c0345f71d24f302f6b0d822063d9f9be

Scopus. (2022b). Citation Overview. https://www.scopus.com/cto2/main.uri?ctoId=CTODS_1521540207&authors=7007061267&origin=AuthorNamesList

Smith, A., & Stirling, A. (2010). *Synthesis, part of a Special Feature on Transitions, Resilience and Governance: Linking Technological, Ecological and Political Systems The Politics of Social-ecological Resilience and Sustainable Socio-technical Transitions* (tech. rep.).

Smith, R. A., & Desrochers, P. R. (2020). Should algorithms be regulated by government? *Canadian Public Administration*, *63*(4), 563–581. https://doi.org/10.1111/capa.12393

Snyder, H. (2019). Literature review as a research methodology: An overview and guidelines. *Journal of Business Research*, *104*, 333–339. https://doi.org/10.1016/j.jbusres.2019.07.039

Splinter-van Kan, H., & Hol, A. (2021). *Eindrapportage Raadspersonen Belastingdienst* (tech. rep.). https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=&ved=2ahUKEwj_r4KzrcH5AhUZg_0HHQ4ABRoQFnoECAYQAQ&url=https%3A%2F%2Fopen.overheid.nl%2Frepository%2Fronl-e501d6c4-9986-4d94-baea-869f962154be%2F1%2Fpdf%2Feindrapportage-raadspersonen-belastingdienst.pdf&usg=AOvVaw0y25V3utKk63qOuFWPNxr9

Steinglass M. (2013). Dutch uproar over Bulgarian benefit fraud. https://www.ft.com/content/7afd3bd6-bcac-11e2-b344-00144feab7de

Tanner, T., Bahadur, A., & Moench, M. (2017). *Shaping policy for development odi.org Challenges for resilience policy and practice* (tech. rep.). www.odi.org/resilience-scan.

Tax Authority. (2020). Herstellen wat fout is gegaan. https://services.belastingdienst.nl/toeslagen-herstel/

Termeer, C. J. A. M., & Dewulf, A. (2019). A small wins framework to overcome the evaluation paradox of governing wicked problems. *Policy and Society*, *38*(2), 298–314. https://doi.org/10.1080/14494035.2018.1497933

Termeer, C. J., Dewulf, A., & Biesbroek, R. (2019). A critical assessment of the wicked problem concept: relevance and usefulness for policy science and practice. https://doi.org/10.1080/14494035.2019.1617971

Turnbull, N., & Hoppe, R. (2019). Problematizing 'wickedness': a critique of the wicked problems concept, from philosophy to practice. *Policy and Society*, *38*(2), 315–337. https://doi.org/10.1080/14494035.2018.1488796

Tweede Kamer. (n.d.). De Nederlandse democratie. https://www.tweedekamer.nl/zo_werkt_de_kamer/de_nederlandse_democratie

Tweede Kamer. (2022). Zoeken in Kamervragen: kinderopvangtoeslag. https://www.tweedekamer.nl/zoeken?fld_prl_kamerstuk=Kamervragen&fld_tk_categorie=Kamerstukken&qry=kinderopvangtoeslag&cfg=tksearch&sta=1&srt=date%3Aasc%3Adate

United Nations. (n.d.). Peace, Justice, and Strong Institutions. https://www.undp.org/sustainable-development-goals#peace-justice-and-strong-institutions

United States government. (n.d.-a). Advancing Trustworthy AI. https://www.ai.gov/strategic-pillars/advancing-trustworthy-ai/#Metrics-Assessment-Tools-and-Technical-Standards-for-AI

United States government. (n.d.-b). Applications. https://www.ai.gov/strategic-pillars/applications/#National-Security-and-Defense

United States government. (n.d.-c). International Cooperation. https://www.ai.gov/strategic-pillars/international-cooperation/

Van Daalen, E. C., Schaffernicht, M., & Mayer, I. (2014). *System Dynamics and Serious Games* (tech. rep.). www.socialimpactgames.com

Van Der Zee, D. J., Holkenborg, B., & Robinson, S. (2012). Conceptual modeling for simulation-based serious gaming. *Decision Support Systems*, *54*(1), 33–45. https://doi.org/10.1016/j.dss.2012.03.006

Van Rij, M. (2022). Reactie op nadere verzoeken Fraude signalering Voorziening. www.rijksoverheid.nl

van Thiel, S., & Yesilkagit, K. (2011). Good neighbours or distant friends?: Trust between Dutch ministries and their executive agencies. *Public Management Review*, *13*(6), 783–802. https://doi.org/10.1080/14719037.2010.539111

Vierhauser, M., Islam, M. N. A., Agrawal, A., Cleland-Huang, J., & Mason, J. (2021). Hazard analysis for human-on-the-loop interactions in sUAS systems. *ESEC/FSE 2021 - Proceedings of the 29th ACM Joint Meeting European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, 8–19. https://doi.org/10.1145/3468264.3468534

vom Brocke, J., Hevner, A., & Maedche, A. (2020). Introduction to Design Science Research. https://doi.org/10.1007/978-3-030-46781-4{_}1

Von Bóné, E. (2019). The Influence of the trias politica of Montesquieu on the first Dutch Constitution. *Comparative law* (pp. 111–121). Routledge. https://doi.org/10.4324/9780429423246-8

Vydra, S., & Klievink, B. (2019). Techno-optimism and policy-pessimism in the public sector big data debate. *Government Information Quarterly*, *36*(4). https://doi.org/10.1016/j.giq.2019.05.010

Walker, G. H., Stanton, N. A., Salmon, P. M., & Jenkins, D. P. (2008). A review of sociotechnical systems theory: A classic concept for new command and control paradigms. *Theoretical Issues in Ergonomics Science*, *9*(6), 479–499. https://doi.org/10.1080/14639220701635470

Walker, W. E., Lempert, R. J., & Kwakkel, J. H. (2013). *Deep Uncertainty Uncertainty In Model-Based Decision Support* (tech. rep.). https://doi.org/10.1007/978-1-4419-1153-7{_}1140

Waterson, P., Robertson, M. M., Cooke, N. J., Militello, L., Roth, E., & Stanton, N. A. (2015). Defining the methodological challenges and opportunities for an effective science of sociotechnical systems and safety. *Ergonomics*, *58*(4), 565–599. https://doi.org/10.1080/00140139.2015.1015622

Weekers F.H.H., & Opstelten I.W. (2013). Wet aanpak fraude toeslagen en fiscaliteit.

Wolford, B. (n.d.). What is GDPR, the EU's new data protection law? https://gdpr.eu/what-is-gdpr/

Wu, P. P. Y., Fookes, C., Pitchforth, J., & Mengersen, K. (2015). A framework for model integration and holistic modelling of socio-technical systems. *Decision Support Systems*, *71*, 14–27. https://doi.org/10.1016/j.dss.2015.01.006

Zhu, X., Singla, A., Zilles, S., & Rafferty, A. N. (2018). An Overview of Machine Teaching. http://arxiv.org/abs/1801.05927

A

Value correlations

This appendix shows the correlations for the values tested in chapter 7. The correlations between all values are depicted in figure A.1. The correlations between values for the final decision-maker (I) and the technological context (C) are illustrated in figure A.2. The correlations between the final decision-maker (I) and the policy execution (B) is shown in figure A.3. Lastly, the correlations between the technological context and the policy execution is presented in figure A.4.

Figure A.1: Correlations between all values

Correlations

		Iefficiency	Iefficacy	Iequality	Iprivacy	Itransparency	Cefficiency	Cefficacy	Cequality	Cprivacy	Ctransparency	Befficiency	Befficacy	Bequality	Bprivacy	Btransparency
Iefficiency	Pearson Correlation	--														
	N	15														
Iefficacy	Pearson Correlation	,578*	--													
	Sig. (2-tailed)	0,024														
	N	15	15													
Iequality	Pearson Correlation	-,706**	-0,377	--												
	Sig. (2-tailed)	0,003	0,166													
	N	15	15	15												
Iprivacy	Pearson Correlation	-0,294	-,581*	-0,035	--											
	Sig. (2-tailed)	0,288	0,023	0,901												
	N	15	15	15	15											
Itransparency	Pearson Correlation	-0,500	-0,392	0,168	-0,322	--										
	Sig. (2-tailed)	0,058	0,148	0,549	0,242											
	N	15	15	15	15	15										
Cefficiency	Pearson Correlation	,562*	0,304	-,518*	-0,400	0,096	--									
	Sig. (2-tailed)	0,029	0,271	0,048	0,139	0,733										
	N	15	15	15	15	15	15									
Cefficacy	Pearson Correlation	0,105	0,427	0,008	-,734**	0,357	,529*	--								
	Sig. (2-tailed)	0,709	0,113	0,977	0,002	0,192	0,043									
	N	15	15	15	15	15	15	15								
Cequality	Pearson Correlation	-0,259	-0,134	,579*	0,400	-,572*	-,725**	-0,492	--							
	Sig. (2-tailed)	0,351	0,633	0,024	0,139	0,026	0,002	0,063								
	N	15	15	15	15	15	15	15	15							
Cprivacy	Pearson Correlation	0,091	-0,357	-0,099	,642**	-0,424	-0,415	-,879**	0,442	--						
	Sig. (2-tailed)	0,746	0,191	0,726	0,010	0,115	0,124	0,000	0,099							
	N	15	15	15	15	15	15	15	15	15						
Ctransparency	Pearson Correlation	-0,424	-0,143	-0,073	-0,087	,678**	-0,156	0,077	-0,472	-0,370	--					
	Sig. (2-tailed)	0,115	0,610	0,796	0,759	0,006	0,580	0,784	0,076	0,175						
	N	15	15	15	15	15	15	15	15	15	15					
Befficiency	Pearson Correlation	0,465	0,493	-0,262	-0,300	-0,287	,599*	0,438	-0,316	-0,343	-0,238	--				
	Sig. (2-tailed)	0,081	0,062	0,345	0,278	0,299	0,018	0,102	0,251	0,211	0,393					
	N	15	15	15	15	15	15	15	15	15	15	15				
Befficacy	Pearson Correlation	-0,018	0,283	0,162	-0,461	0,160	0,481	,664**	-0,296	-,655**	-0,027	,649**	--			
	Sig. (2-tailed)	0,950	0,307	0,565	0,084	0,568	0,070	0,007	0,285	0,008	0,925	0,009				
	N	15	15	15	15	15	15	15	15	15	15	15	15			
Bequality	Pearson Correlation	-0,225	-0,045	0,232	0,255	-0,224	-,742**	-,518*	,658**	0,456	-0,062	-,750**	-,699**	--		
	Sig. (2-tailed)	0,421	0,874	0,404	0,359	0,421	0,002	0,048	0,008	0,087	0,825	0,001	0,004			
	N	15	15	15	15	15	15	15	15	15	15	15	15	15		
Bprivacy	Pearson Correlation	0,010	-0,466	0,045	,548*	-0,278	-0,381	-,718**	0,503	,781**	-0,384	-,531*	-,788**	,550*	--	
	Sig. (2-tailed)	0,973	0,080	0,874	0,034	0,316	0,162	0,003	0,056	0,001	0,157	0,042	0,000	0,034		
	N	15	15	15	15	15	15	15	15	15	15	15	15	15	15	
Btransparency	Pearson Correlation	-0,216	-0,136	-0,210	-0,143	,649**	0,080	0,248	-,628*	-0,370	,798**	-0,280	-0,066	-0,151	-0,413	--
	Sig. (2-tailed)	0,440	0,629	0,454	0,610	0,009	0,776	0,372	0,012	0,174	0,000	0,311	0,816	0,591	0,126	
	N	15	15	15	15	15	15	15	15	15	15	15	15	15	15	15

*. Correlation is significant at the 0.05 level (2-tailed).

**. Correlation is significant at the 0.01 level (2-tailed).

Figure A.2: Correlations between the final decision maker's values and for the colleague

Correlations

Control Variables			Iefficiency	Iefficacy	Iequality	Iprivacy	Itransparency	Cefficiency	Cefficacy	Cequality	Cprivacy	Ctransparency
Befficiency & Befficacy & Bequality & Bprivacy & Btransparency	Iefficiency	Correlation	1,000	0,711	-0,677	-0,473	-0,349	0,656	0,226	-0,447	0,007	-0,493
		Significance		0,021	0,031	0,167	0,322	0,039	0,529	0,195	0,985	0,148
		df	0	8	8	8	8	8	8	8	8	8
	Iefficacy	Correlation	0,711	1,000	-0,576	-0,576	-0,171	0,604	0,408	-0,511	-0,256	-0,280
		Significance	0,021		0,082	0,081	0,636	0,064	0,241	0,131	0,475	0,434
		df	8	0	8	8	8	8	8	8	8	8
	Iequality	Correlation	-0,677	-0,576	1,000	-0,018	0,231	-0,584	0,025	0,646	-0,163	0,128
		Significance	0,031	0,082		0,961	0,522	0,076	0,945	0,044	0,652	0,724
		df	8	8	0	8	8	8	8	8	8	8
	Iprivacy	Correlation	-0,473	-0,576	-0,018	1,000	-0,450	-0,412	-0,616	0,475	0,437	0,137
		Significance	0,167	0,081	0,961		0,192	0,236	0,058	0,165	0,207	0,705
		df	8	8	8	0	8	8	8	8	8	8
	Itransparency	Correlation	-0,349	-0,171	0,231	-0,450	1,000	-0,031	0,302	-0,467	-0,289	0,487
		Significance	0,322	0,636	0,522	0,192		0,933	0,397	0,173	0,417	0,153
		df	8	8	8	8	0	8	8	8	8	8
	Cefficiency	Correlation	0,656	0,604	-0,584	-0,412	-0,031	1,000	0,382	-0,710	-0,278	-0,443
		Significance	0,039	0,064	0,076	0,236	0,933		0,276	0,021	0,436	0,200
		df	8	8	8	8	8	0	8	8	8	8
	Cefficacy	Correlation	0,226	0,408	0,025	-0,616	0,302	0,382	1,000	-0,277	-0,742	-0,371
		Significance	0,529	0,241	0,945	0,058	0,397	0,276		0,439	0,014	0,291
		df	8	8	8	8	8	8	0	8	8	8
	Cequality	Correlation	-0,447	-0,511	0,646	0,475	-0,467	-0,710	-0,277	1,000	0,079	-0,033
		Significance	0,195	0,131	0,044	0,165	0,173	0,021	0,439		0,828	0,927
		df	8	8	8	8	8	8	8	0	8	8
	Cprivacy	Correlation	0,007	-0,256	-0,163	0,437	-0,289	-0,278	-0,742	0,079	1,000	-0,102
		Significance	0,985	0,475	0,652	0,207	0,417	0,436	0,014	0,828		0,778
		df	8	8	8	8	8	8	8	8	0	8
	Ctransparency	Correlation	-0,493	-0,280	0,128	0,137	0,487	-0,443	-0,371	-0,033	-0,102	1,000
		Significance	0,148	0,434	0,724	0,705	0,153	0,200	0,291	0,927	0,778	
		df	8	8	8	8	8	8	8	8	8	0

Figure A.3: Correlations between the final decision maker's values and for the boss

Correlations

Control Variables			Iefficiency	Iefficacy	Iequality	Iprivacy	Itransparency	Befficiency	Befficacy	Bequality	Bprivacy	Btransparency
Cefficiency & Cefficacy & Cequality & Cprivacy & Ctransparency	Iefficiency	Correlation	1,000	0,704	-0,726	-0,438	-0,703	0,225	-0,308	0,257	-0,090	-0,011
		Significance		0,023	0,017	0,205	0,023	0,533	0,387	0,473	0,805	0,975
		df	0	8	8	8	8	8	8	8	8	8
	Iefficacy	Correlation	0,704	1,000	-0,654	-0,462	-0,745	0,340	-0,076	0,356	-0,385	-0,197
		Significance	0,023			0,179	0,013	0,336	0,835	0,312	0,271	0,586
		df	8	0	8	8	8	8	8	8	8	8
	Iequality	Correlation	-0,726	-0,654	1,000	-0,122	0,757	-0,137	0,336	-0,344	0,045	0,014
		Significance	0,017	0,040		0,736	0,011	0,706	0,343	0,330	0,901	0,969
		df	8	8	0	8	8	8	8	8	8	8
	Iprivacy	Correlation	-0,438	-0,462	-0,122	1,000	-0,074	0,036	0,041	-0,325	0,034	0,213
		Significance	0,205	0,179	0,736		0,839	0,921	0,911	0,360	0,925	0,555
		df	8	8	8	0	8	8	8	8	8	8
	Itransparency	Correlation	-0,703	-0,745	0,757	-0,074	1,000	-0,546	0,067	0,021	0,448	-0,034
		Significance	0,023	0,013	0,011	0,839		0,103	0,854	0,955	0,195	0,927
		df	8	8	8	8	0	8	8	8	8	8
	Befficiency	Correlation	0,225	0,340	-0,137	0,036	-0,546	1,000	0,498	-0,653	-0,582	-0,520
		Significance	0,533	0,336	0,706	0,921	0,103		0,143	0,041	0,078	0,124
		df	8	8	8	8	8	0	8	8	8	8
	Befficacy	Correlation	-0,308	-0,076	0,336	0,041	0,067	0,498	1,000	-0,658	-0,720	-0,367
		Significance	0,387	0,835	0,343	0,911	0,854	0,143		0,039	0,019	0,297
		df	8	8	8	8	8	8	0	8	8	8
	Bequality	Correlation	0,257	0,356	-0,344	-0,325	0,021	-0,653	-0,658	1,000	0,343	0,232
		Significance	0,473	0,312	0,330	0,360	0,955	0,041	0,039		0,331	0,519
		df	8	8	8	8	8	8	8	0	8	8
	Bprivacy	Correlation	-0,090	-0,385	0,045	0,034	0,448	-0,582	-0,720	0,343	1,000	0,006
		Significance	0,805	0,271	0,901	0,925	0,195	0,078	0,019	0,331		0,987
		df	8	8	8	8	8	8	8	8	0	8
	Btransparency	Correlation	-0,011	-0,197	0,014	0,213	-0,034	-0,520	-0,367	0,232	0,006	1,000
		Significance	0,975	0,586	0,969	0,555	0,927	0,124	0,297	0,519	0,987	
		df	8	8	8	8	8	8	8	8	8	0

Figure A.4: Correlations for the colleague's and boss decisions, from the perspective of the final decision-maker

Correlations

Control Variables			Cefficiency	Cefficacy	Cequality	Cprivacy	Ctransparency	Befficiency	Befficacy	Bequality	Bprivacy	Btransparency
lefficiency & lefficacy & lequality & lprivacy & ltransparency	Cefficiency	Correlation	1,000	0,562	-0,752	-0,564	-0,429	0,610	0,680	-0,718	-0,454	-0,189
		Significance		0,091	0,012	0,089	0,216	0,061	0,031	0,019	0,188	0,602
		df	0	8	8	8	8	8	8	8	8	8
	Cefficacy	Correlation	0,562	1,000	-0,456	-0,777	-0,364	0,476	0,550	-0,574	-0,513	0,057
		Significance	0,091		0,186	0,008	0,302	0,164	0,100	0,083	0,130	0,875
		df	8	0	8	8	8	8	8	8	8	8
	Cequality	Correlation	-0,752	-0,456	1,000	0,438	0,009	-0,737	-0,614	0,834	0,685	-0,217
		Significance	0,012	0,186		0,205	0,981	0,015	0,059	0,003	0,029	0,547
		df	8	8	0	8	8	8	8	8	8	8
	Cprivacy	Correlation	-0,564	-0,777	0,438	1,000	-0,126	-0,443	-0,532	0,561	0,592	-0,214
		Significance	0,089	0,008	0,205		0,729	0,200	0,113	0,091	0,071	0,552
		df	8	8	8	0	8	8	8	8	8	8
	Ctransparency	Correlation	-0,429	-0,364	0,009	-0,126	1,000	-0,067	-0,206	0,070	-0,211	0,577
		Significance	0,216	0,302	0,981	0,729		0,855	0,568	0,849	0,558	0,081
		df	8	8	8	8	0	8	8	8	8	8
	Befficiency	Correlation	0,610	0,476	-0,737	-0,443	-0,067	1,000	0,752	-0,912	-0,648	-0,244
		Significance	0,061	0,164	0,015	0,200	0,855		0,012	0,000	0,043	0,497
		df	8	8	8	8	8	0	8	8	8	8
	Befficacy	Correlation	0,680	0,550	-0,614	-0,532	-0,206	0,752	1,000	-0,844	-0,790	-0,205
		Significance	0,031	0,100	0,059	0,113	0,568	0,012		0,002	0,007	0,571
		df	8	8	8	8	8	8	0	8	8	8
	Bequality	Correlation	-0,718	-0,574	0,834	0,561	0,070	-0,912	-0,844	1,000	0,726	0,059
		Significance	0,019	0,083	0,003	0,091	0,849	0,000	0,002		0,017	0,872
		df	8	8	8	8	8	8	8	0	8	8
	Bprivacy	Correlation	-0,454	-0,513	0,685	0,592	-0,211	-0,648	-0,790	0,726	1,000	-0,361
		Significance	0,188	0,130	0,029	0,071	0,558	0,043	0,007	0,017		0,306
		df	8	8	8	8	8	8	8	8	0	8
	Btransparency	Correlation	-0,189	0,057	-0,217	-0,214	0,577	-0,244	-0,205	0,059	-0,361	1,000
		Significance	0,602	0,875	0,547	0,552	0,081	0,497	0,571	0,872	0,306	
		df	8	8	8	8	8	8	8	8	8	0