AI Governance in the City of Amsterdam

City of Amsterdam

TU Delft

AI GOVERNANCE IN THE CITY OF AMSTERDAM

Scrutinising Vulnerabilities of Public Sector AI Systems

A thesis submitted to the Delft University of Technology in partial fulfillment of the requirements for the degree of

Master of Science in Engineering and Policy Analysis

by

Daniël Brom

4369424

To be defended in public on July 1 2021

Graduation committee

Chair	Hans de Bruijn	Multi-Actor Systems
Daily Supervisor	Roel Dobbe	Engineering Systems and Services
First Supervisor	Haiko van der Voort	Multi-Actor Systems
Second Supervisor	Sem Nouws	Engineering Systems and Services
External Supervisor	Linda van de Fliert	City of Amsterdam
External Supervisor	Anne-Maartje Douqué	City of Amsterdam

Preface

Here it is, my master's thesis. After months of hard work, I hope to deliver a research project which gives useful insights into the vulnerabilities of working with AI in the public sector, and the governance trade-offs that occur when the City of Amsterdam tries to address these vulnerabilities. If not, I hope that you at least find some joy in reading the report. If even that is too much to ask, I am happy to talk to you about this project in real life, now the C-word pandemic seems to finally come to an end.

Since this preface probably is as close to an Oscars speech as I'll ever get, I would like to thank some people. First of all, Linda and Anne-Maartje, who allowed me to join the City of Amsterdam for a research internship and for really letting me become a part of the CTO and Public Tech Team. And Linda and Selma, thank you for the fun and helpful weekly supervision meetings too. Unfortunately they always had to be held online. From TU Delft, I would like to thank Roel and Sem who took the time to help me out every week, and got me to think about context-dependent vulnerabilities as a focal point in this thesis. Hans, for always being critical in a very constructive way and telling me when I am producing bureaucratic governance nonsense. I really enjoyed working on my humble contributions to your book on institutions for privacy as well. Lastly I want to say thank you to Haiko, your feedback during the fixed thesis-meetings was always very helpful to structure my research and come up with a consistent storyline.

Now let's get emotional. No, kidding. Cheers to Hadas, my family and friends. I am very happy looking back over the past seven years, having met so many great people during my time as a student and having had a great time in my student houses, in Berlin, at Curius, Ariston, at Rijkswaterstaat, the City of Amsterdam, TPM, and other places.

Das war einmal. Now let's look ahead!

All the best,

Daniël Brom
Delft, June 17, 2021

Executive Summary

Scandals in which governmental ADM tools played a role, have recently brought about political and societal debate about the potential harms to citizens that such automated systems potentially bring. In the *Toeslagenaffaire*, the Dutch Tax and Customs Administration used an algorithmic risk classification system for fraud detection with childcare benefit applications (Parlementaire Ondervragingscommissie Kinderopvangtoeslag, 2020). Until 2018, nationality of the parents was one of the indicators to check for fraud. The Dutch Data Protection Authority (DPA) concludes in its 2020 investigation report that using nationality as a fraud indicator was unnecessary, discriminatory, and unlawful with respect to the EU General Data Protection Regulation (Autoriteit Persoonsgegevens, 2020). In Amsterdam, law firm SOLV raises questions about the lawfulness and effectiveness of a municipal ADM for detection of Airbnb fraud (van Dorp, 2020).

This thesis focuses on ADM systems which contain an AI component. For AI, algorithms are the means to create a system of computational codes with human-related competencies like perception, understanding, and action (Mnih et al., 2015; Wirtz, Weyerer, & Geyer, 2019). However, the AI in this thesis, as in most applications, concerns so-called Narrow AI (NAI). Narrow refers to the goal-focused and specialised characteristics of the solutions. Such AI technologies show no general intelligence, but intelligence in a specific area (Pennachin & Goertzel, 2007). AI is explored in the context of process governance according to Pierre & Peters (2020): governance as steering and coordinating during the process of AI development. Governments seeking to design governance which safeguards citizens from potential harms, encounter challenges inherent to AI, for example lacking understanding with policymakers due to technological complexity (Wirtz et al., 2019). It is also well-known that designing governance strategies for public sector AISs leads to value trade-offs, for example when decision-makers have to find a balance between privacy protection and accuracy of the system (Dobbe & Raji, 2019).

Despite the societal urgency of the subject and the large amounts of scientific articles on public sector AI governance, some knowledge gaps still have to be addressed. Zuiderwijk et al. (2021, p.16) call for *multidisciplinary* studies which *develop and test theories* about AI governance *specifically for the public sector*. The preliminary literature review also reveals a lacking understanding of how potential citizens harms due to AI result in governance requirements for governments, as the link between vulnerabilities of AISs and governance requirements is rarely made. Lastly, keeping up with practical development of AI in governments worldwide is necessary for public sector AI governance scholarship. Currently there is a lack of empirical research on governmental AISs (Zuiderwijk et al., 2021). Based on the knowledge gaps perceived, the main research question for this study is:

In public sector AI systems, what are emerging vulnerabilities for citizens and how do these translate into governance requirements for decision-makers?

The approach to answer this question is an adjusted form of Theory Building from Case Study as proposed by Eisenhardt (1989). The adjustment is to do create theory upfront, to be tested during the case study, where it is common not to do so. The theoretical model based on an integrative literature review is validated and improved with empirical insights from semi-structured interviews, with case study actors from three cases: Reporting issues in public space, Illegal holiday rental housing risk, and Automated parking control. After having assessed the theoretical model, case study insights also serve to find governance requirements from the CoA practice which can be linked to the model. But first, the integrative literature review helps to better understand what the implications of AI use in the public sector are.

Three main types of reasons to use AI for governmental operations are: improving efficiency of decision-making, improving effectiveness of decision-making and inherent advantages it brings to citizens. But AISs bring disadvantages too, for example in the form of biased decision-making or potential corruption of automated systems. The introductory literature review also revealed several types of mitigation measures for inconveniences of AISs: technological measures, data measures, monitoring and evaluation measures, legal measures, organisational measures, and user engagement and citizen agency. The overview of disadvantages point to directions for the theoretical framework of AIS vulnerabilities.

A layered 'onion model' presents four relevant contexts to consider for AIS vulnerabilities specifically in the public sector: *AI model*, *Model deployment*, *Political administrative*, and *Societal*. If a government decides to buy or develop some AIS, the first source of potential vulnerabilities is the model itself. There is a department which deploys the model for its daily operations, where other problems may occur. Then there are the political-administrative overarching policy objectives which are involved with everything the specific government does. Lastly, the societal context represents the actors who are ultimately affected by the AIS, namely citizens, where other vulnerabilities appear. The contextual layers together constitute a unity which covers the scope of every governmental AIS. Vulnerabilities differ from "Overfitting and underfitting" in the AI model context to "Negative impact on workforce" in the model deployment context to "Reward hacking" in the societal context. These examples illustrate the different nature of vulnerabilities: e.g. technological, socioeconomic and behavioural. Vulnerabilities emerge in different stages of the model development, but this research distinguishes two main phases: either before or after implementation.

The expert interviews for the three case studies in total yields 52 case-specific vulnerabilities of the AISs. 6 of them did not fall within one of the vulnerabilities in the model contexts. This is an acceptably low share and the model is considered to be valid for further use. The vulnerabilities that were not interpretable using the model, point towards three complements: *model being adopted without enough capacity to control it*, *inadequacies in the requisite security of the AI model* and *unwanted strategic behaviour with the AI model*. From the vulnerabilities that were found, several lessons are learned. One: the AIS of Reporting issues in public space creates new forms of biases in favour of certain subgroups of citizens who are e.g. more outspoken or have higher trust in government. The involved actors did not acknowledge that it then becomes a political choice to what extent the AIS must influence the municipal strategy regarding public space management. Two: as seen in the Automated parking control case, shifting discretionary power becomes a striking vulnerability when an important share of the AIS is outsourced to an external party. Questions were raised here e.g. about whether the current KPIs in the contract between Egis Parking Services and the CoA lead to fair routing of automated scanning cars for Amsterdam's citizens. Three: based on the Illegal holiday rental housing risk vulnerabilities: for fraud detection systems based on large amounts of data from different sources, it is questionable whether governments

can find legal foundations or domain-knowledge arguments for all relationships the AI model creates to come up with risk profiles.

The case study results demonstrate that dealing with vulnerabilities in one of the four model contexts often complicates dealing with vulnerabilities in the other contexts. Hence, four governance requirements dilemmas found in the CoA practice which relate to the vulnerabilities model are:
I-a Increase the impact of the AI model on the decision-making process to ensure its added value and create balance with the downsides of the AI system. vs
I-b Decrease the impact of the AI model on the decision-making process until the effects of model errors are acceptable or errors are still possible to mitigate.
II-a Leave model developers with enough time and professional freedom to create high-end technological products. vs
II-b Ensure that developers create models which are explainable and functional for governmental employees to be used in their daily practice.
III-a Be transparent about the AI systems you use. Communicate actively and provide opportunities for citizens' idea contribution and participation. vs
III-b Foster objective representations of reality by your AI systems and prevent new forms of bias caused by citizen participation.
IV-a Stimulate innovation with regards to AI development within your organisation and do not restrict every innovative project upfront. vs
IV-b Ensure that AI development projects have a clear and proportional contribution to an agreed policy objective.

Reflecting on all results, it is not the existence of AIS vulnerabilities, but merely the absence of a thorough process to settle considerations about the vulnerabilities which is the current challenge for governments. If governments use reports like this to understand the context-specific and interdependent vulnerabilities, this forms a first step towards this process and creating room for politically responsible decision-makers to make the relevant trade-offs. The next step is to document such considerations and create overviews of best practices, so that learning between development teams and their managers within or outside the CoA and other governmental bodies is allowed. Several policies would contribute to establishing this maturity in AI governance:
i) Goals for the AIS may develop over time, but to at least have a shared belief in what the system is ought to do and keep discussing this over the lifecycle of the system, helps to find out about the relevant consideration of vulnerabilities as just described. Besides having an agreed purpose for the AIS, the distinction between efficiency goals, effectiveness goals, citizen well-being goals and innovative goals should always be made clearly as well.
ii) Resolving challenges in one context, often seems to lead to emergence of vulnerabilities in others. Understanding such tensions and trade-offs should be the primary focus when assessing the risks of using ADM within governments. Using the vulnerabilities model can be of help to do so: multidisciplinary teams with actors from all contexts can use the model for directions to think about the vulnerabilities they encounter.
iii) Although only being transparent is not enough to deal with AIS vulnerabilities, it is essential and thus highly recommendable to provide societal actors with more opportunities to get involved with governmental use of AI. Media, interpreted as societal actors, are then enabled to take up their role as well.
iv) Decision-makers must analyse the bias that occurs *by using the AI* rather than the *bias of the AI itself*. This is oftentimes not so much the problem, as in the cases Reporting issues in public space and Automated parking control, or it is the exact reason for the algorithmic system to be used in the first place, as in the Illegal housing rental case. Only focusing on mitigation of the algorithmic bias itself would therefore miss the point.

Contents

List of acronyms

ADM Algorithmic decision-making
AGI Artificial General Intelligence
AI Artificial Intelligence
AIS AI System
BD Big data
CTO Chief Technology Office
CoA City of Amsterdam - *Gemeente Amsterdam*
DDDM Data-driven decision-making
DL Deep Learning
DPA Dutch Data Protection Authority - *Autoriteit Persoonsgegevens*
EPA Engineering & Policy Analysis - *Delft University of Technology MSc Program*
FBI Federal Bureau of Investigation
IA Intelligent Agent
ML Machine Learning
NAI Narrow Artificial Intelligence
NN Neural Networks
RL Reinforcement Learning
SAVRY Structured Assessment of Violence and Risk in Youth
SL Supervised Learning
SyRI System Risk Indication - *Systeem Risico Indicatie*
UL Unsupervised Learning
XAI Explainable AI

List of Figures

List of Tables

Chapter 1

Introduction

Artificial Intelligence (AI) is a vexed subject in societal, political and scientific debates. For individual citizens, the question rises what potential harms for them emerge if their governments adopt AI-fueled decision-making methods. Such potential harms also became apparent in different cases of AI-tools in the City of Amsterdam (CoA). This thesis report, conducted for the TU Delft MSc program Engineering & Policy Analysis (EPA) and a research internship at the CoA Public Tech Team, deep-dives into the subject of potential harms for individual caused by governmental AI use. First, this chapter introduces the research context, main research concepts, and the main research question based on existing knowledge gaps.

1.1 Context: AI in the Netherlands and Amsterdam

By the time of writing, January 2021, the Netherlands is still under the spell of an institutional crisis. This crisis lead to the resignation of the Dutch cabinet on January 15, 2021. AI played an important role in this scandal. The CoA enjoys a keen societal and media interest for its deployed AI systems as well. This section sketches the research context of this thesis about public sector use of AI. It does so to illustrate the potential harms AI can cause to citizens, as well as to emphasise the urgency and relevancy of research on good governance for public sector AI.

1.1.1 AI's questionable role in a disastrous system

On December 17, the Parliamentary Interrogation Committee Childcare Allowances presented its tellingly titled report: "Unprecedented Injustice". From 2014 until 2019, tens of thousands of parents who receive childcare benefits from the Dutch Tax and Customs Administration, are unduly regarded as fraudulent. Small mistakes made by parents lead to fraud classifications by the Administration. Accused parents are forced to pay back enormous amounts of money, regularly without a payment scheme being offered. This leads to profound damaging consequences for these parents, oftentimes having to sell their house, falling into large debts, and ending up in years of poverty. The scandal is called "Benefits Affair". The Parliamentary Interrogation Committee concludes that behind this harsh approach lies a political urgency for efficiency and strict governmental fraud prevention systems in general (Parlementaire Ondervragingscommissie Kinderopvangtoeslag, 2020)[1].

[1] Report only available in Dutch

To detect potential fraudulence at an early stage, the Dutch Tax and Customs Administration uses an algorithmic risk classification system for childcare benefit applications. This automated and self-learning system ranks the potential of childcare benefit fraud based on application forms filled out by parents (Parlementaire Ondervragingscommissie Kinderopvangtoeslag, 2020). Applications that receive a high risk score from the system, are double-checked by a Tax and Customs Administration official. The system uses dozens of indicators for its risk rankings. Until 2018, nationality of the parents was one of these indicators. This means that parents with a non-Dutch nationality had a higher chance of being thoroughly checked. Combined with conclusions about the unduly classification of fraudulent cases, this means parents with a dual nationality had a higher chance of enormous financial claims, leading to the previously mentioned problems. The Dutch Data Protection Authority (DPA) concludes in its 2020 investigation report that using nationality as a fraud indicator was unnecessary, discriminatory, and unlawful with respect to the EU General Data Protection Regulation (Autoriteit Persoonsgegevens, 2020).

Using AI and algorithms for predictive and profiling purposes seems inevitable for organisations in modern societies. Human decisions are replaced by algorithmic decision-making (ADM) at rapid pace (Kroll et al., 2017). AI potentially contributes to governmental institutions' performance by increasing effectiveness and efficiency. But oftentimes, the drawbacks of using AI by governmental institutions are given too little attention (Janssen & Kuk, 2016). Wrong use of personal citizen information by ADM systems can lead to discrimination and other ethical concerns (Autoriteit Persoonsgegevens, 2020; Corbett-Davies & Goel, 2018). The Benefits Affair is another example which points towards a persistent issue for governmental institutions, namely problematic information management (Parlementaire Ondervragingscommissie Kinderopvangtoeslag, 2020). It sharply demonstrates the disastrous effects wrong governmental use of AI systems can have for particular citizen subgroups in the Netherlands. Since AI systems have increasing impacts on citizens, their fairness becomes increasingly important as well (Corbett-Davies & Goel, 2018).

1.1.2 Controversy in Amsterdam

Amsterdam-based law firm SOLV directs attention to the municipality's algorithm for detection of Airbnb fraud cases. SOLV's Van Dorp (2020) raises questions about the lawfulness and effectiveness of such a fraud detection system. The adoption of the fraud detection system by Amsterdam is an extra sensitive topic. In February 2020, the Dutch The Hague Court found another fraud detection AI system, System Risk Indication (SyRI), to be violating article 8 of the European Convention on Human Rights (Linders, 2020). The Court ruled SyRI to underperform on many aspects: accountability, transparency, privacy, prevention of discrimination and stigmatisation, and lack of possibilities for civilian opposition procedures (Huisman, 2020). This leaves the question whether the CoA is able to protect its citizens from these severe consequences when they deploy their own Airbnb fraud detection system (van Dorp, 2020).

The quality of the output of AI systems raises concern in the media as well. De Ruijter (2021) writes about the high number of objections to parking fines in Amsterdam, which are often handed out by automated parking control systems on cars. Out of 551,150 fines by scanning cars, 70,513 were objected to in 2019. 66% of these objections had to be accepted by the municipality.

Taken together, the CoA does not escape questions about the protection of their citizens from AI induced risks. In doing so, they are no exception to other local, national, and international governments all over the world. There are growing concerns

about the risks of deploying more and more algorithmic and data-driven AI systems. Governmental organisations like the CoA must pursue their citizens' best interests by definition. Therefore, AI deployment in public contexts like those of Amsterdam, require special attention.

1.2 Introducing AI and governance of public sector AI

This section outlines definitions of AI and interrelated concepts, as well as an introduction of what is now known about governance of public sector AI.

1.2.1 AI, algorithms, data-driven decision-making, and big data

Scientific articles on AI, algorithms, data-driven decision-making (DDDM) and big data (BD) seem to use these terms interchangeably, albeit they do have different meanings and connotations. This section presents the meanings of these concepts, as well as how they are interrelated, as interpreted for this research. This provides clarity for the rest of this report.

Big data, data-driven decision-making, and algorithms

Although the term BD implies an emphasis on the amount of data processed, the data set size is certainly not the most important factor. BD relates to new capacities empowered by large-scale data gathering and processing techniques. Scholars therefore tend to interpret BD as new developments in how data is used, not the changes in the data sets itself (Klievink, Romijn, Cunningham, & de Bruijn, 2017). Early scholars in the field Boyd & Crawford (2012) describe how the term stands at the intersection of three developments:

- *Technology*: increasing computing power
- *Analytical methods*: the ability to identify patterns by cross-referencing data
- *Belief*: the claim that insights to patterns in large data sets lead to new knowledge

Using algorithms, meaning is provided to BD. Simply put, algorithms process BD. The concepts are therefore not to be analysed separately if one studies decision-making based on BD (Janssen & Kuk, 2016; Hill, 2016). Algorithms in the light of Boyd & Crawford's (2012) pioneering BD understanding, are made possible by *Technology*, are mathematical rules and therefore *Analytical Methods* in essence, and used because of *Belief*. Put shortly, an algorithm is some sort of data-fueled set of mathematical rules which is believed to provide knowledge to the organisation. These mathematical rules can differ strongly in terms of complexity.

DDDM, often referred to in slightly other terms like data-informed decision-making, is an overarching concept. It enables data scientists and decision-makers to make public decisions based on BD (Van Der Voort, Klievink, Arnaboldi, & Meijer, 2019). As follows from this section, the BD-component implies the connection to algorithms as well.

Focus of this research: Artificial Intelligence

AI, which includes learning methods for data processing but also takes many other forms, refers to intelligent systems that update themselves. The study of AI is to study intelligent agents (IAs) that pursue some defined goal. The IA contains an AI model which represents the environment of the agent mathematically. The IA furthermore has an objective function, which is the mathematical form of the agent's goal in its environment. In its environment, the IA perceives information through its sensors and

acts by means of its outputs, which are decisions or 'actuators' (Dobbe, Gilbert, & Mintz, 2021).

AI is furthermore characterised by the human-related competencies of perception, understanding and action (Wirtz et al., 2019). Intelligent combinations of algorithms are applied for the creation of understanding and beliefs about real-world complex systems in light of BD (Ghahramani, 2015; Lu, Li, Chen, Kim, & Serikawa, 2018). For AI, algorithms are the means to create a system of computational codes with its own competencies and challenge-resolving skills (Mnih et al., 2015). Seen from the DDDM definition in section 1.2.1, AI used for governmental decision-making is a tool for DDDM. Vice versa, DDDM systems often contain AI components. Including the terms BD-driven and algorithms in the scope of AI for this research leaves open the opportunity to gain important insights from work on BD and ADM in general, rather than turning a blind eye and to only focus on works which use the term AI.

AI includes many forms of models and technologies. Amongst them are Neural Networks (NN), Robotics, and Computer Vision (Pennachin & Goertzel, 2007). This thesis' goal is not to dive into all AI techniques too deeply. Mentioning these types of models serves to illustrate the wide range of possible technological solutions which would all be considered AI models. One particular subset of AI techniques needs further explanation: Machine Learning (ML). ML models are an often used class of models for DDDM by public sector organisations, using administrative data. ML model structures identify relevant patterns in different data sets to turn them into usable information for DDDM (Veale & Brass, 2019). ML techniques, which are a subset of AI techniques, contains other subsets of techniques in itself as well. For example, NN are sets of algorithms used to recreate neural learning processes. They can process data structures composed of multiple arrays, like images or spoken words. These NN hugely profit from increasing computational power and available data, and therefore fuel so-called Deep Learning (DL) engineering techniques (Lecun, Bengio, & Hinton, 2015). NN and DL are just some of the few computer science subsets which would all be considered ML and therefore AI.

As the AI models considered for this research, which the CoA uses, are all ML models, a distinction of the three types of ML needs to be made based on (Simplilearn, 2020; Lison, 2012). The first one is Supervised Learning (SL). SL models are trained to learn a task based on labeled data: the labels represents the right outcome, and the model is then trained to predict the right labels/outputs if it is fed with new input variables from large data sets, which are called features. Unsupervised Learning (UL) models take unlabeled data as input and are trained to learn about underlying patterns in the data to find features themselves and predict the right outputs. Lastly, Reinforcement Learning (RL) models are trained based on rewards in a trial-and-error way of learning. An agent, often a robot or game playing computer, is rewarded for the output it delivers based on some evaluation of outputs.

Artificial Intelligence is actually narrow intelligence

This short subsection seeks to dispel some confusion about AI. The term implies generic intelligence from the computer system, but at this moment, the systems used by public sector organisations actually show no generic intelligence. If one does not recognises this, it would be easy to overestimate both the possibilities and dangers of using AI.

Pennachin & Goertzelf (2007) describe how most contemporary research on AI is focused on so-called Narrow AI (NAI). The same holds for this thesis. Narrow refers to the goal-focused and specialised characteristics of the solutions. NAI systems are trained to perform single tasks. They show no general intelligence, but intelligence

in a specific area. The performed single tasks can be relatively simple, like playing a game of chess, but more complex single tasks fall within the definition of NAI too: as long as they are singular, goal-focused and specialised. In a subtle way, NAI changed the meaning of intelligence in its computer science context. It has shifted away from a general consciousness and understanding of the world to the ability of performing particular tasks in a more efficient or effective way. The notion of intelligence has entered stage due to the fact that people assumed the need of some form of human intelligence to perform these particular tasks. Nevertheless, NAI solutions can deliver high performances, but the techniques do not allow for actual new and strategic knowledge generation by integrating several knowledge sources in cutting-edge ways. Experts expect Artificial General Intelligence (AGI) to be decades away from now, in their modest estimations. Narayanan (2019) expects even these estimations to be fiercely optimistic. Since AGI is a long way off, this thesis for simplicity reasons uses AI instead of NAI.

1.2.2 AI system perspective

In scientific articles and reports about AI, one often encounters the term 'AI system' (AIS), for example in (Morley et al., 2021; AI HLEG, 2019). However, an explanation of what an AIS is and why it is preferable to use the system perspective rather than just simply AI, often lacks. This section first explains why the (sociotechnical) system perspective is useful and then specifically defines what an AIS is.

De Bruijn & Herder (2009) describe the analysis of sociotechnical systems. Sociotechnical systems, which the deployment of AI in public sector operations clearly belongs to, first have technological subsystems. If one of the subsystems does not function as expected, it is doubtless that the functioning of other subsystems is influenced as well. For example, in an AI technological context: if the data gathering of some prediction model is inadequate, the AI model will have less predictive power. Apart from the technical system perspective, the involved actors need to be analysed as well. This is the networked multi-actor component. Combining these perspectives will lead to more intelligent designs of the sociotechnical AI systems that are able to "stand the test of real-world implementation" (de Bruijn & Herder, 2009, p.991).

So to take the system perspective is to broaden the scope of analysis. Not only the AI model itself is of interest, but the interdependencies between sociotechnical subsystems and involved actors must be taken into account as well. But what is an AIS? Krafft et al. (2020) find that AI researchers tend to overemphasize the technological ideal thinking of AI in their definitions, whereas policymakers assign too much weight to AI's ability to reproduce human capabilities. Sensible AIS definitions therefore combine both worlds, with current and future AI technologies taking center stage, but not being the only matter of research. As in the paper by Kraft et al. (2020), the OECD definition of an AIS is considered to incorporate all important elements of analysis for this thesis:

"AI system: An AI system is a machine-based system that can, for a given set of human-defined objectives, make predictions, recommendations, or decisions influencing real or virtual environments. AI systems are designed to operate with varying levels of autonomy." (Yeung, 2020)

Some last important notions about AISs deserve attention. Due to the automated and self-learning aspects of AI technologies, combined with the complex public environments in which they are considered for this research, AISs show emergent behavior. The behavior of the AIS can be unexpected or unintended with regard to its human-defined objectives (de Bruijn & Herder, 2009). A last aspect is that public AISs either automate,

aid, or replace human decision-making in the public sector. By doing so, the public AISs impact human welfare by definition (Richardson, Chan, et al., 2019).

1.2.3 AI governance

This research places AISs in the context of governance, as protecting citizens from AI-inspirited harms is a governance question for the CoA and other governments. This section introduces the current concerns about AI governance as found in the literature, preceded by a clarification on a suitable scoping of governance.

Scoping 'governance' for this research

Governance is a highly generic term which needs scoping to be workable. If AISs in Amsterdam indeed bring about harms to citizens, this is a collective problem which asks for collective addressing by a governmental organisation. A governance perspective which directs attention to the organisation of a governmental body, as well as its relationship with its citizens, is useful in this case. Pierre & Pieters (2020) are influential authors in the governance research field. They present multiple structural lenses on governance: governance-as-hierarchies, governance-as-markets, governance-as-networks. Structural perspectives on governance do not fit this research best, because structural perspectives particularly consider institutions. This research merely focuses on the process of AI development and its outcomes in the form of citizen safety risks. Enter the lens of *process governance*: **Governance as steering and coordinating**.

This perspective allows for the assumption that government steers society (Pierre & Peters, 2020), which is a necessary assumption for this research on damaging consequences of AI deployment. After all, harms to citizens are only established if the AISs have significant impact on citizen behaviour, individual rights or lives. This significant impact is not exclusively caused by the municipality's legal powers, but also settled in its control over critical resources and its pursuit for their role of collective interest Gestalt (Pierre & Peters, 2020). Defining governance in terms of steering and coordinating allows for a research focus on the objectives of steering ('who is affected by the AIS'), what collective interest should be pursued ('What are rightful trade-offs for the AIS') and the dynamics of the multi-actor environment in Amsterdam which produces the outcomes.

1.2.4 Governance challenges for public sector AI

Governments seeking to design governance which safeguards citizens from potential harms caused by public sector AISs, encounter governance challenges inherent to the use of AI. This section introduces some of these topics to provide an idea of what decision-makers from the CoA may run into in their daily practice.

A first concern which is shared amongst many authors is accountability in AISs. Practices in DDDM must be made accountable to safeguard objective judgement (Ziewitz, 2016). Transparency and opaqueness of algorithmic or AISs is often problematic and could hinder public accountability. At the same time, it is questionable whether opening of the algorithmic black box leads to increased transparency or only allows malicious gaming and exploitation of the systems (Janssen & Kuk, 2016). An important aspect of accountability is the issue of intent. Accidental and deliberate misuse of AI are hard to distinguish (Floridi et al., 2018).

Technological complexity of algorithmic code, especially when there are multiple contributors to the code, disallows complete understanding of involved actors. According to Wirtz, Weyerer & Geyer (2019), AI making use of large amounts of data leads to

diffusion of information which is increasingly hard to decipher. Ziewitz (2016, p.7) calls this difficulty of understanding "inscrutability". Oftentimes there is not enough available capacity of AI knowledge within public organisations (Wirtz et al., 2019). This complexity hinders effective governance measures for AISs (Janssen & Kuk, 2016). Differences between the knowledge base of both analysts and decision-makers may results in scepticism about the ADM system on both sides. Designers who visualise and present insights can play an important mediating role here (Van Der Voort et al., 2019). Citizens as well lack understanding of how public organisations make decisions based on complex AI. This can either be caused by the complexity of the used technology or by the abundance of system components and data variables (Algemene Rekenkamer, 2021).

Another theme in AI governance is human agency and oversight. Citron & Pasquale (2014), amongst others, use the distinction of Human-in-the-Loop, Human-on-the-Loop and Human-out-of-the-Loop. Floridi et al. (2018, p.698) call for "meta-autonomy": human decision-makers should always retain the power to override decisions and determine which organisational decisions should be human-made. Danaher (2016) argues that no decisions with zero human influence can be legitimate, as human participation is essential for legitimacy. In current literature, Meaningful Human Control is a more conventional term than Human-in-the-loop to describe the role of human agency and oversight in deployment of autonomous systems. Verdiesen, de Sio & Dignum (2020) perceive a lack of clear understanding of what Meaningful Human Control should look like in practice. Nevertheless, they do point towards accountability, moral responsibility, controllability and oversight as important building blocks.

1.2.5 Public sector AI governance leads to trade-offs

Section 1.2.4 presents particular AIS governance challenges to give a preliminary idea of the research matter. Another research topic became apparent during the preliminary literature review of public sector AI governance as well. Designing governance measures for AISs and ADM inherently leads to trade-offs. These trade-offs are interesting because only to focus on singular governance issues would oversimplify the complexity of AIS governance. Relevant trade-offs are important for the contributions of this thesis, since they illustrate the necessity of choosing values over others in this research area. An introductory overview of trade-offs is presented in this section.

Liability or litigation of human domain-experts for ADM, stimulates experts to critically assess the decisions proposed by AISs. At the same time, strict ultimate responsibilities may stimulate experts to not use the AI solutions, scrutinizing the effectiveness of the AI application. This illustrates a trade-off between expert tacit knowledge, self-esteem and responsibility, versus the flourishing - if designed carefully - of the AIS (Cohen, Amarasingham, Shah, Xie, & Lo, 2014). The possibilities for effective AI use are also at stake when many data sources which potentially violate privacy laws are eliminated from the system. There is a general balance between protecting privacy and accuracy of the system (Dobbe & Raji, 2019).

Sometimes, AI leads to tensions between protecting consumer rights by companies and governmental goals. For example, Apple was reluctant to permit the US' Federal Bureau of Investigation (FBI) access to their iPhone encryption, which the FBI needed for a criminal investigation. This lead to a trial, which the FBI won (Janssen & Kuk, 2016). This anecdote is related to a more generic AI governance trade-off: choosing between the personal experience and protection, or increasing utility for the whole population (Cohen et al., 2014).

Chowdhury & Sloane (2020) also pose a tension between private interests and democratic control. Private AISs partly capture the public digital infrastructure, and in that

way lead to important public task fulfillment without democratic control. Governmental organisations cannot simply rely on the company's right intentions, arguments or craftsmanship, but nevertheless are hindered to conduct risks analysis or controls due to the intellectual property (Algemene Rekenkamer, 2021).

Trade-offs between AI fairness means, both reasonable, are described by Whittaker et al. (2018). If developers choose not to include sensitive attributes for the data input of the system, which potentially prevents bias or discrimination, this withholds the developers to conduct post hoc mathematical tests for independence of sensitive attributes. Kleinberg et al. (2017) additionally argue that there is an inherent trade-off between the accuracy of AI predictions and the equal distribution of prediction errors over citizen subgroups.

A trade-off between optimal learning for the AIS and maximum effort to protect citizens' well-being is emergent as well. Take for example the domain of health care. If one would search for optimal training of surgery AISs, testing on real patients is the best simulation. At the same time, this involves taking the risk of for example testing cuts for these patients. Wirtz et al. (2019) call this the link between AI advancement and protection of humankind.

1.3 Knowledge Gaps and Research Question

The Benefits Affair and controversy in Amsterdam, presented in section 1.1, make clear that public sector AI use does not only cause positive effects for citizens. The CoA and other governments seek for governance strategies which adequately address the potential harms caused to citizens by using AI for decision-making. Section 1.2.4 and section 1.2.5 treat some of the governance challenges and governance trade-offs with respect to AI. But what scientific knowledge of downsides of public sector AI and governance strategies to deal with those lack? This section presents the scientific knowledge gaps addressed in this research.

First of all, theory development for public sector AI is understudied. In their up-to-date research agenda for public sector AI governance, based on a literature review of 26 carefully selected studies, Zuiderwijk et al. (2021, p.16) call this "under-theorization" of the AI public governance scholarship. They call for multidisciplinary theoretical foundations and studies which develop, test or extend theories about AI in the public sector. In doing so, the context-dependent aspect of public sector AI risk governance is important, as AI governance should be progressive to a level "proportional to the risk level associated with a specific combination of technology and context" (Morley et al., 2021, p.15). This context-dependency of scholarship on AI governance in the public sector oftentimes lacks.

This thesis aims at developing and testing such theory with insights from multiple scientific disciplines. The specific form of this theory development is a framework which focuses on vulnerabilities of public AISs, which contributes to the explanation of how vulnerabilities emerge. There are both technological and governance reasons for problems to emerge. Currently there is a lack of frameworks concerning the vulnerabilities of using AI by governments. To construct this public sector AIS vulnerabilities theory also addresses another existing knowledge gap: most conceptual frameworks in the field of AI governance and its risks are not specific to the public sector (Zuiderwijk et al., 2021).

During the preparatory literature study to introduce the research topic, a next knowledge gap became clear. Scholars in the AI governance field seem to focus on the downsides of using ADM in general, as well as AI governance challenges and trade-offs.

But a clear link between the potential harms to citizens from public sector AISs on the one side, and requirements in terms of governance strategies for public sector decision-makers on the other side lacks. The abovementioned theory of emerging vulnerabilities in public sector AIS will aid in establishing this link.

Furthermore, Zuiderwijk et al. (2021) call for investigation of best practices concerning public sector governance strategies to deal with risks of AISs. Keeping up with practical development of AI in governments worldwide is necessary for public sector AI governance scholarship. Otherwise, the lack of evidence-based and contextual research potentially leads to failures in public AI governance. This thesis seeks to do focus on best practices in vulnerability mitigation by conducting interviews with practitioners from the CoA. Answers to questions about current methods to deal with vulnerabilities serve as a means to find such best practices. Working with empirical data to assess and complement the theoretical framework helps to bridge the existing knowledge gap of empirical research on the impact of AISs adoption for governments (Zuiderwijk et al., 2021). Lastly, the constructed theoretical framework of this thesis addresses the context-dependency of public sector AI vulnerabilities for citizens.

Building on the scientific knowledge gap, the research question for this thesis is:

Main Research Question

In public sector AI systems, what are emerging vulnerabilities for citizens and how do these translate into governance requirements for decision-makers?

AI systems (AISs) are defined in section 1.2.2. Chapter 4 provides clarity on the meaning of vulnerabilities, whereas chapter 6 elaborates on the definition of governance requirements. This means that only the definition of *emerging* remains:

The vulnerabilities which occur when governments like the CoA use AI, result from interaction effects of technological difficulties, interaction effects of steering and co-ordination of the AISs and, most important, combinations of such technological and governance aspects.

1.4 Thesis Outline

This report first presents the research approach in chapter 2. The literature review is set out in chapter 3 and chapter 4, where the first concerns generic information about AI in the public sector and the latter presents the conceptual model. Then, chapter 5 gives an overview of the case study results. Chapter 6 contains AI governance requirements and measures from the CoA practice. The research is capped with the discussion and conclusion in chapter 7 and chapter 8 respectively.

Chapter 2

Research Approach and Methods

This chapter presents all relevant information for the approach to bridge the knowledge gap and answer the main research question for this thesis:

In public sector AI systems, what are emerging vulnerabilities for citizens and how do these translate into governance requirements for decision-makers?

2.1 Type of Research: Exploratory

The nature of this research is exploratory. This thesis' goal is to develop pertinent hypotheses and theoretical constructs which can be used for further research, which suits the characterisation of exploratory research by Yin (1994). Public sector AI is a relatively new field of research, as mentioned earlier this chapter. In such a new research field, an exploratory case study is useful to identify theoretical ideas which can be used for further research (Yin, 2018).

During the conduction of this research, there is an opportunity to be an intern at the CoA. This internship takes place at the Public Tech Team of the Chief Technology Office (CTO). Doing the internship brings several benefits for this research. First, it provides the opportunity to easily contact involved stakeholders from the municipality, to do interviews and reflect on ideas. Furthermore, feedback from people who work with AI in their daily public sector practice can be provided. Lastly, joining municipal meetings and conducting research whilst participating in the real-life practice contributes to a nuanced view on reality (Flyvbjerg, 2006).

2.2 Sub-questions for This Research

This section presents the sub-questions for this research as well as the reasoning behind them.

To understand why governments like the CoA use AISs in the first place, one must also know what uses AI models can have for governmental organisations. Furthermore, governments are not completely handed down to AIS's downsides, as they can also conduct measures for control or prevention of the AI induced risks. In short, we need a nuanced view on public sector AI. As presented in chapter 1, AISs in the public sector can have negative consequences. But anecdotal arguments are not enough to

understand this new development in the public sector. This knowledge then leads to an understanding of what remains unknown, which lays the groundwork for the theoretical construct of this research. The first two sub-questions concern these knowledge gaps:

Sub-question 1:
What are the advantages and disadvantages of using AI models in the public sector and what types of mitigation measures are available to deal with the disadvantages?

Sub-question 2:
Which knowledge of these disadvantages is lacking and what kind of theoretical construct would contribute to this knowledge?

If governmental organisations want to use AI and better be able to mitigate the downsides, only understanding the general concerns with AI is not enough. After having answered sub-question 2, it becomes clear what a relevant theoretical model to understand vulnerabilities must contain. The next question is, seen from the relevant perspectives of a public organisation like the CoA, where vulnerabilities of AISs originate from. The third sub-question addresses this knowledge gap:

Sub-question 3:
Seen from an AI system perspective, from which relevant contexts do vulnerabilities originate and how can these be structured in a conceptual model?

Next a check for validity of the theoretical construct is important because this model is the main theoretical contribution of this thesis, and theory building from case study is often criticised for lacking validity. Therefore the first uses of the case study results are to validate and complement the theoretical construct, which reflects in sub-question 4:

Sub-question 4:
What do the case-specific vulnerabilities found in the City of Amsterdam's practice say about the validity and directions for complementing of the conceptual vulnerabilities model?

Because there is an existing knowledge gap of best practices for good governance strategies concerning the risks of public sector AI - see section 1.3 - the governance requirements based on the vulnerabilities model are derived from practice rather than from literature. These serve to answer sub-question 5:

Sub-question 5:
What are relevant governance requirements and mitigation measures, found during conduction of the case studies, to deal with context-dependent vulnerabilities of public sector AI systems?

2.3 Research Method per Sub-question

Two different research methods are required to answer the sub-questions. This section presents these methods in detail, as well as the tools to enable the methods.

Sub-question 1, 2 & 3: Integrative literature review

Integrative literature review is an appropriate literature study method for the first three sub-questions. Integrative literature review is a way of assessing literature that allows for synthesising, in order to create an author's own conceptualisation and perspective on the research topic (Torraco, 2005). The qualitative data is not required to be fully comprehensive. Rather, the purpose of an integrative literature review is not to cover all data, but to select the useful and combine different perspectives (Snyder, 2019). So the theoretical model seeks to create a structured understanding of AIS vulnerabilities, but this method allows freedom to only select the useful sources for ones own combinations and interpretations. This means that no exact search queries are necessary.

A drawback of Integrative literature review is the difficulty to provide transparency on the exact search methods and integration of theories (Snyder, 2019). The exact search methods are not relevant since no search queries are used. Instead, the earlier mentioned discussions with involved CoA workers, CoA supervisors and TU Delft supervisors serve to set directions for the literature review and interpretation. To provide transparency, the literature review results will include an overview table of reviewed literature and an overview table of preliminary discussions with the involved CoA workers. The discussions are not recorded and no empirical data are gathered from the meetings, as they only serve as part of the iterative process of structuring and finding literature to construct a conceptual model.

The literature is reviewed with desk research, making use of scientific literature found through Scopus (preferably), and complemented with Google Scholar. Relevant reports, newspaper articles, lecture slides and webpages are used as well, as long as they contribute to answering the sub-questions. A standard analysis tool is unnecessary, the concepts and definitions are interpreted and intertwined by the researcher.

Sub-questions 4 & 5: Semi-structured expert interviews

Expert interviews are a useful method to gather data for exploratory research in a concentrated way. The promise of unproblematic access to relevant data makes it an often used research method in social empirical research (Bogner, Littig, & Menz, 2009). It is important not to be seen as a layperson by the interviewees, to have meaningful conversations and to not be educated (Bogner & Menz, 2009). The preliminary literature review for sub-question 1 aids to overcome this problem. By having an idea of the research variables, one can competently enter the assessment of expert knowledge (Pfadenhauer, 2009).

This specific interview category leaves space for ad hoc elaboration by not using the same list of questions for every interviewee. At the same time it ensures an overarching thematic approach, by determining equal question *categories* for all interviewees (Qu & Dumay, 2011). It is also referred to as open and topic-guided at the same time (Meuser & Nagel, 2009). A semi-structured interview often ensures the enclosure of hidden organisational and human behavior which is relevant to the interviewer. A major drawback of this particular technique is the large amount of effort it requires in advance. A general disadvantage of interviews in scientific research is the dilemma of whether to disclose the intent of the interview - to build trust and clarification - or not, to prevent clouding the interviewee's responses (Qu & Dumay, 2011). It was decided to do disclose the intent, because this was already clear for most interviewed people

due to internal communications and it seemingly did not hinder interviewees to be as open as possible.

2.3.1 Data gathering and processing

The answers to the interview questions form the data for this method, so the answers have to be recorded and processed in a structured way. Requirements are independence of different interviewee's answers and a reasonable amount of interviews, approximately eight to fifteen, so the answers can be processed within the given time for this thesis. The number of interviews conducted in the end is nine. Based on the literature review for sub-questions 1 and 2, the most relevant interviewees are selected. The interviewed experts are presented in the results sections of this thesis.

All recorded interviews are transcribed. It is impossible to interpret all transcriptions in a reasonable way by just reading them over and over. Therefore, coding based on Auerbach & Silverstein (2003) is used. Coding allows to overcome this impossibility and enable theory building from the transcripts. Coding also brings validity because the theoretical construct is supported by the coded data. The steps that need to be undertaken are: select relevant text, find repeating ideas and themes, generalise to theoretical constructs and then end up with your own theoretical narrative which potentially is a new research concern to others. Atlas.ti 9[1], which is available trough TU Delft channels, is used as a software tool for convenient coding.

2.4 Research Stages

This research consists of four subsequent stages, answering the five sub-questions presented in section 2.2. These together answer the main research question. The first three research stages are **descriptive**, whereas the last stage is **prescriptive** research.

2.4.1 Stage 1: Introduce public sector AI, understand what we do understand, and understand what we do not understand

Several intermediate stages are necessary to answer the main research question. Based on the thorough and influential description of theory building from case study research by Eisenhardt (1989), this thesis first takes the following research step:

- Construct ex ante ideas about the research topic variables based on existing literature, but avoid taking conclusions about causal relationships.

A nuanced view is necessary to be well-prepared for the expert interviews: the interviews will have less value if the interviewees have to explain basal aspects of the AIS they are involved with.

The scientific knowledge gap identified in section 1.3 calls for development of theoretical constructs on public sector AI, as well as attention for context-specific research. But it remains unknown what exactly would be a relevant theoretical construct to develop, and what the relevant contexts are. This part of the thesis presents a first overview of disadvantages of public sector AI use. Based on this preliminary overview, the two unknowns as just described are addressed. The found disadvantages also set directions to search for literature to build the theoretical construct.

Lastly, another step contributes to stage 1 of this research. Preliminary conversations with different municipality workers from the CoA take place to serve the goals as

[1] https://atlasti.com/

described in this research stage: understand the AISs to prepare for the interviews, as well as get an idea of what theoretical construct would contribute to knowledge that is now lacking. This creates an iterative process of discussing AISs with actors from the field, search for relevant literature based on these discussions, discuss first findings from literature with other actors, again searching for relevant literature.

2.4.2 Stage 2: Building a relevant theoretical construct to understand the origin of public sector AI vulnerabilities

After having answered the first two sub-questions, a first understanding of the benefits, downsides and potential mitigation measures is available. These aspects are relevant for a nuanced answering of the main research question. Furthermore, a first risk overview and discussions with CoA employees reveal what knowledge is still lacking about how AI vulnerabilities originate from different perspectives of the AI as a system. Constructing the conceptual theoretical model, to address these vulnerabilities, remains an iterative process. Discussions with CoA workers, CoA supervisors, and TU Delft supervisors set directives for the literature search and construction of the model.

After having created preliminary ideas and a conceptual model in the first two stages, the research enters its next stage. This is where the data collection for actual case study starts. The next steps are again based on (Eisenhardt, 1989):

- 'Enter the field': start observing and collecting information by as many means possible
- Cross-compare case information and observations in a structured way
- Present the case study results from different data sources in different ways. Use tabular displays, transcripts, quotes, quantitative summaries.
- Adjust the initial theoretical constructs, research steps and possibly methods based on the first case study results
- Supplement theoretical findings with emergent themes and insights from case studies
- Verify theoretical contributions by comparing the designed framework of risk categorisation with case evidence.
- Present theoretical contributions based on case study results, cross-case comparisons and iterative comparison of existing literature and own findings.

2.4.3 Stage 3: Understanding the cases, validate and complement the vulnerabilities model

To add to existing knowledge from the literature, bottom-up knowledge from experience with downsides of public sector AISs in Amsterdam is relevant. Besides, these practitioners' insights are also useful to validate the conceptual model from the previous research stage. If the findings from three different cases to a large extent fit into the generic conceptual model, this would confirm the validity of the conceptual model of vulnerabilities.

2.4.4 Stage 4: Finding governance requirements and strategies to deal with vulnerabilities

Stage 4 is the prescriptive stage of this research. Now the conceptual model is validated and directives for complementing it are provided, it can be used for further analysis. If vulnerabilities emerge in different contexts, public sector organisations face requirements to deal with these vulnerabilities in different contexts as well. Mitigation

measures support the CoA when facing these governance requirements to deal with vulnerabilities.

Based on the experiences from CoA workers as discussed during interviews, the relevant governance requirements to deal with vulnerabilities in different AIS contexts are presented. Furthermore, interesting and relevant mitigation strategies from the CoA's practice, categorised based on the mitigation categories from this research' phase 1, are presented as well.

2.5 General Research Approach

Based on Eisenhardt (1989), the research approach for this thesis is **Theory Building from Case Study**. The execution however is somewhat adjusted. Eisenhardt (1989) recommends not to create theory before case analysis. But the identified knowledge gaps in section 1.3 reveal that public sector AI governance is "under-theorized" and empirical research on the assessment and evaluation of AI-related disadvantages for citizens is still uncommon. This thesis' adjusted form of Theory building from case study combines the best of both worlds to address these gaps. Preliminary discussions with CoA employees provide ideas for theory development and set directions for literature review, the wide available set of literature on AI governance and AI risks lays the groundwork for the theoretical construct. Then the theoretical contribution is validated and improved with empirical insights to ensure its applicability in public sector reality and to allow the interpretation of cases regarding governance requirements.

Key concerns for this approach are whether it is possible to validate your theoretical constructs, as well as loosing sight of simplicity and coming up with overly detailed and narrow theory (Eisenhardt, 1989). This thesis deals with these concerns in the following way. To make sure that the validity of the conceptual model can be tested for, the conceptual model is divided into several contextual layers, which are subdivided into actual vulnerabilities for public sector AISs. In this way, the interviews can serve as validation method by comparing the mentioned downsides of AISs from the Amsterdam practice with the vulnerabilities from the conceptual model. To immediately use case insights for validation makes sure that the theoretical construct from case study is already validated once, which addresses the criticism of hard-to-validate constructs from case study. By stressing the fact that the focal point of the vulnerabilities model is the context-dependent interpretation and the filled-in vulnerabilities are just some of the vulnerabilities found in these layers, the model is not overly narrow and can be complemented if necessary.

2.5.1 Elaboration on the Case Study

Multiple case study is an appropriate research method if no case exists which explains all relevant phenomena for theory building - e.g. a critical, extreme, or unique case. None of such cases is available for the AISs deployed by the CoA. Multiple case study then allows for better confirmation or rejection of theory, highly robust results and finding policy implications through cross-case comparison (Yin, 1994).

Another choice is to either focus on different sub-units of analysis or perceive the case study in a holistic manner. Sub-units are different responsible officials as well as different organisational layers (Yin, 1994). To understand the vulnerabilities of AISs it is particularly interesting to study different perspectives on these risks and relationships between different departments of the municipality, which all contribute to the dynamics of the AIS. Combining the facts that multiple sub-units of analysis are

involved and more than one case is studied, this research is a **multiple, embedded case study**.

Table 2.1 presents a short description of the three AISs from the CoA used as cases for this study. The descriptions are based on the information provided in the municipality's Algorithm Register (City of Amsterdam, n.d.).

Table 2.1: Brief description of three AISs the municipality deploys

Department	AI System	Short Description
Economic services	Automated parking control	Cars equipped with cameras scan license plates from parked cars. License plates are identified and checked for parking rights if the car owner e.g. did not pay a parking fee, a ticket is issued. Human parking inspector checks scans for validity and special circumstances.
Economic services	Illegal holiday rental housing risk	Based on fraud cases of the past years, personal information, and spatial information, this AI system supports officials with the investigation of potential fraud notifications filed by citizens. The system calculates the probability of fraud and advises officials whether or not to further investigate the notification.
City management	Reporting issues in public space	Traffic problems, rubbish and other disturbances can be reported to the municipality. Notifications have to be categorised in order for the right department to be able to resolve the issue. This AI system recognises the right category for citizen notifications, in order to prevent delays due to wrong manual categorisation.

The three systems all have their own relevant unique characteristics, which are the conditions to be added to the research portfolio according to Yin (1994). These are the added values per case:

Illegal holiday rental housing risk strikes because of the sensitive data that are necessary for the system to function. The system also takes over an important task of CoA officials, the prioritising of potential fraud notifications. The impact on whom is checked and potentially penalised could therefore potentially be high.

Reporting issues in public space is interesting because of the potential of biased decision-making, as preliminary talks with CoA workers revealed. Due to linguistic or other errors the system sometimes does not process notifications correctly. This probably leads to disadvantages for citizen subgroups.

Automated parking control is not developed by the City of Amsterdam's own developers, unlike the other three cases. The system is designed by Egis Parking Services. In early talks with municipality workers, it appeared that there might be some tension due to the external development. Public servants might not have a satisfactory grip on the way these external developers fulfill their job and commercial interests of findings as much illegally parked cars potentially play a role as well. This makes the case interesting in light of citizen risks.

Eisenhardt (1989) strongly recommends to conduct case study research with multiple researchers. Due to master thesis restrictions this is impossible, which might jeopardise the validity of the outcomes. The low number of selected case studies as a result of time limitations, might pose threats to the complexity of designed constructs as well. Therefore, regular feedback sessions with both TU Delft and Amsterdam municipality experts are held. These discussions with academic and public sector experts also partly overcome the problem of conducting the case study alone.

2.5.2 Research flow and structure of the report

The logical flow of answering the sub-questions one by one is presented in figure 2.1. The boxes contain the chapters which serve to report about the different research stages and answer the corresponding sub-questions. The research methods belonging

to these phases are located bottom left of the dashed larger boxes. The arrows present the outputs of research stages which serve as inputs for subsequent stages. Combining the insights from all different research stages leads to a synthesis and answering of the main research question.

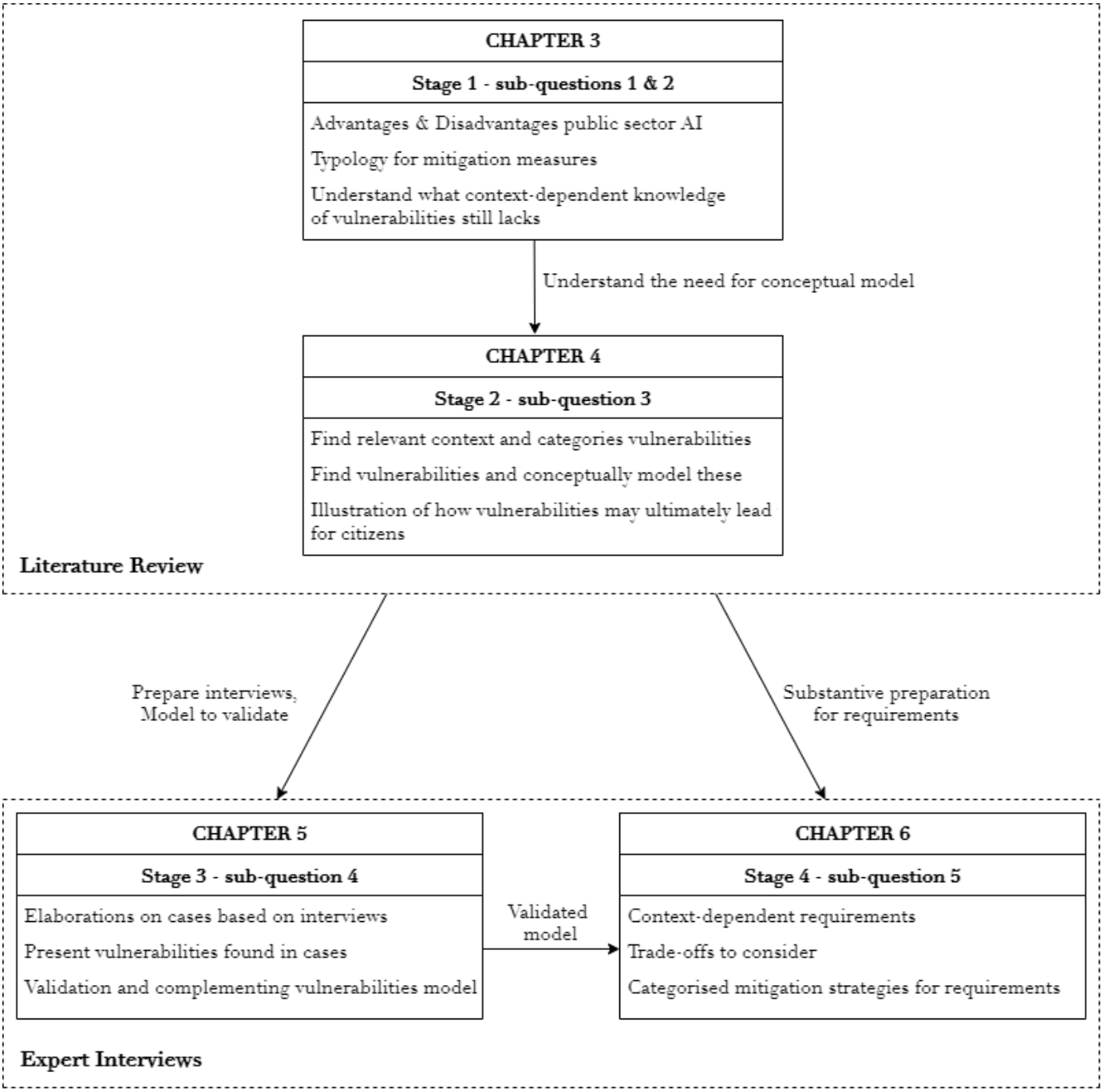

Figure 2.1: Research flow diagram for this research

Chapter 3

Understanding the Use of AI by Public Sector Organisations

This chapter contains the results from stage 1 of this research: "Introduce public sector AI, understand what we do understand, and understand what we do not understand". This first part of the literature review focuses on three elements of this understanding. First, an overview of what reasons public organisations have to use AI is presented. Then, a first overview of downsides which potentially result from adopting AISs in the public sector follows. This preliminary overview serves to understand what knowledge about vulnerabilities of these systems still lacks. Lastly, an elementary indication of different kinds of mitigation measures as found in the (semi-)literature is presented. This indication later helps to categorise and understand the mitigation measures found in the CoA's practice.

3.1 Advantages of using AI in the public sector

The reviewed literature presents two main types of merits for public sector organisations to us AISs in its operations. These are advantages for the government as an organisation, which wants to function properly, and advantages for the citizens governments seek to serve. A subdivision in the advantages to the government as an organisation appears: benefits in terms of efficiency and benefits in terms of effectiveness. Figure 3.1 visualises this categorisation of public sector AI advantages. But what do increased efficiency and effectiveness in a public sector context mean? And what benefits to citizens can digitisation of a governmental process bring to citizens? This subsection elaborates on these main reasons to use AI in public contexts.

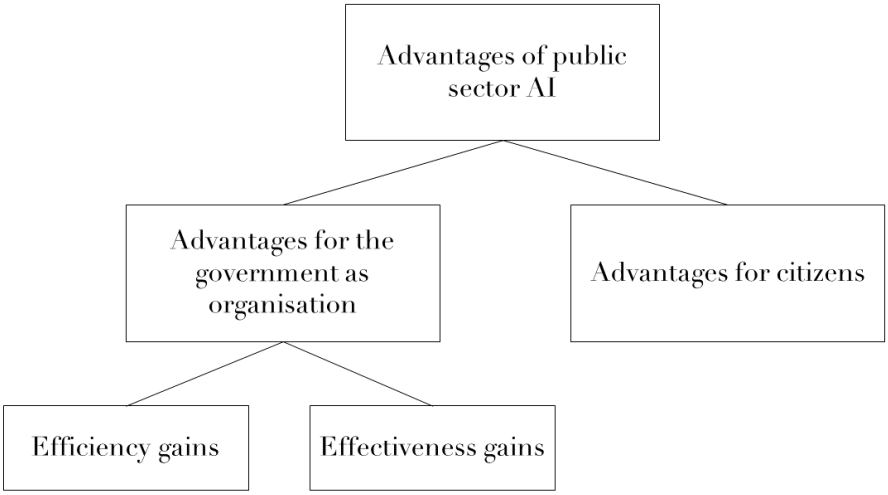

Figure 3.1: Overview of public sector AI advantages

3.1.1 Advantages for the government as an organisation

This section presents two main categories of advantages for organisations like the CoA, which are not inherent benefits for the citizens it should serve, but merely advantages for the organisation itself. This notion is no value judgement but a way to reasonably understand the different kinds of benefits of public sector AI.

Increasing operational efficiency

The coming specific benefits found in (semi-)literature are interpreted as benefits concerning efficiency goals for governments.

Transmit complex or dangerous tasks to machines
Wirtz et al. (2019) mention the possibility for AI technologies to replace humans in conducting dangerous tasks. This saves operational capacity in terms of potential injuries and fatalities as well as thorough time and resources which are necessary for performing dangerous tasks. An example of such a dangerous task which can be replaced by an AIS is found in the safety domain. Firefighters in Europe were allowed to make use of the "SmokeBot". This is an EU-funded project with a robot that is able to navigate through low-visibility environments, mostly places that are on fire. This AIS thus assists firefighters who now do not have to enter a building on fire, risking their lives (Fan, Bennetts, Schaffernicht, & Lilienthal, 2019).

Addition to and better utilisation of existing human capacity
If AISs are used to help conducting a governmental body's daily operations and service processes (Algemene Rekenkamer, 2021), this means that AI supplements operational capacity without having to hire new employees. Within the same time, the governmental organisation conducts more operations and thus increases its efficiency. Besides possibilities to complement human operational capacity, some AI technologies also accommodate to better take advantage of the already existing human capacity. ADM systems for categorisation and prioritisation assist agencies to better aim the human controllers and other resources for their checks and inspections (Wirtz et al., 2019).

Fewer administrative burdens
Veale & Brass (2019) perceive ML systems to have two operational purposes. The first is automation of systems. Automation of systems means that computational power allows a public organisation to use administrative data for models which increase the quantity of performed operational decisions in the same amount of time. In other words, ML systems lead to increased efficiency. Automation systems only can perform social tasks with minor complexity, because of the fact that the outcome of the system must be straightforward and possible to assess in a relatively objective way. This relates to the notion of narrow AI as outlined in section 1.2.1.

If processes are digitised and automated by making use of AI applications, the administrative burden on public organisations can be decreased. The organisations assigns repeatable, logical and rule-based duties to AISs, for example for case workers in human services (Eggers, Fishman, & Kishnani, 2017). When human agents spend less time on administrative tasks, they can perform more others and in this way AI increases public sector efficiency (Wirtz et al., 2019).

Cost savings
If the government organisations save time due to automation of repetitive tasks and lowering administrative burdens, they are able to cut their budget for personnel costs or other non-human resources. Especially for governmental organisations, which are often strapped for financial assets, this can be an important benefit of using AI for its operations (Chowdhury & Sloane, 2020).

Increasing quality and effectiveness of decisions

This category contains technological solutions which not only accelerate the amount of decisions made, but also improve the quality of decisions and hence increase public sector operational effectiveness. But to conclude that the quality of decision-making is improving, is to determine that the ends of public decision-making are quantifiable, a "highly value-laden task in itself" (Veale & Brass, 2019, p.6). The benefits found in the (semi-)literature which are presented here try to grasp what 'increasing effectiveness' or 'higher quality decisions' mean.

Process larger amounts of information in times where much data is available
As described in section 1.2.1, the current BD-era brings many opportunities for the use of large amounts of information. But governments are potentially overwhelmed by the large data amounts of all kinds of types as well (Milakovich, 2012). Since governmental organisations possess and create vast amounts of information about e.g. public space, socio-economics and health, BD holds huge potential for the public sector. But uncertainty about how to process the information and what value it brings to the organisation, may emerge (Klievink et al., 2017). It is therefore sensible to deploy AISs, which can aid in the BD processing and deliver for example management reports and descriptive analyses (Wirtz et al., 2019).

Knowledge standardisation
Wirtz et al. (Wirtz et al., 2019) find that AI in public organisations may lead to 'standardized knowledge': analysis, distribution and sharing of organisational expert knowledge is eased by the AI adoption. If more decision-makers have access to higher quality knowledge, the quality and hence effectiveness of public decisions increases.

Less administrative burden also leads to higher quality of decisions
The same administrative enlightenment which leads to potential efficiency increasing, may also lead to increased quality of decisions or service provision in public contexts. If human service providers need less time for administrative tasks, they save time for

critical tasks and accordingly improve their work of delivering services to the public (Eggers et al., 2017).

Learning systems outperform humans with respect to pattern recognition
If some service provision or task performance of governments is primarily based on pattern recognition, AI solutions may fit these tasks best. Like some AI system A is better at playing chess than human agent B, public AISs can be better at pattern recognition in public contexts. This especially holds for ML systems with cognitive functions that allow the system to learn and respond to new circumstances (Wirtz et al., 2019).

Citizens are more accepting to decisions due to feelings of objectiveness
With policy issues becoming more complex, public decision-makers need to convince the public of their systems to be evidence-based in order to gain public acceptance of their decisions and policies (Kolkman, Campo, Balke-Visser, & Gilbert, 2016). If decision-makers use AISs and in the public opinion, this contributes to evidence based decision-making, using AI results in higher acceptance of government decisions by citizens and hence increases effectiveness of these decisions.

3.1.2 Benefits for citizens relative to conventional decision-making tools

The previous sections presented advantages for the governmental organisations. But these organisations are ought to serve citizens. This last category of advantages found in the (semi-)literature contains the advantages of public sector AISs which are inherent benefits for citizens, rather than advantages for the organisations which use these systems.

Potentially more objective and consistent decisions without human partiality
Human decision-makers sometimes go out on the wrong side of their beds too. They have personal preferences, backgrounds, and biases. Adoption of AI support in decision-making may have value for citizens by making them independent from bureaucrats' prejudices and moods (van Eck, Bovens, & Zouridis, 2018). Due to the computational rule-based learning and decision-making of ML based systems, ADM in the public spheres can also be more consistent, which has intrinsic social benefits (Hind et al., 2019). And even if the ADM system has biases too, which it often does, there might be intrinsic value if AISs provide citizens with the feeling of being treated objectively (Chowdhury & Sloane, 2020).

Citizen expects government to digitise
The internet plays a large role in almost every aspect of societies, businesses and people's daily lives. This fact creates expectations about public service provision and decision-making as well. People expect a digital government. If governmental organisations are digitised, citizens can expect them to enable them for online filling out of forms, which saves them time, for example (Janssen, Hartog, Matheus, Yi Ding, & Kuk, 2020). This societal digitisation demand partly translates into a public wish for ADM and AI applications.

Decreased waiting time for governmental decisions
For some public decisions and services, it is important for citizens to receive the verdict as soon as possible. For example in immigration services. Chun (2007) acknowledges that an immigration agency for application form examination proves to significantly reduce turnaround time. These time savings are one of the reasons that lead the agency to be awarded for better service to citizens.

Possibility to be more transparent to citizens than before

Unlike the human brain with its implicit decision rules, algorithms can be completely open and accessible for interested citizens. The code, decision rules and used data of the AIS can be published for citizens. Although there are problems to overcome with such XAI, decision-making based on AI does bring the potential of increasing transparency for the public sector (Algemene Rekenkamer, 2021).

Potential for better governmental communication with citizens
AISs also bring possibilities for improved communication between citizens and governmental organisations. If citizens are better informed about decisions, they may have higher acceptance towards them or better understand what is expected from them, leading to fewer frictions between public authorities and civilians. Androutsopoulou et al. (2019) describe how the adoption of chatbots, i.e. three different AI technologies (natural language processing, ML and data mining technologies) lead to better citizen-government communication.

Personalised decisions
AI tools provide governments with the opportunity to align their service provision with the citizen's individual needs (Sousa, Melo, Bermejo, Farias, & Gomes, 2019). Services which better fit individual compulsions of citizens will result in higher quality of services. An example from Finland. The Ministry of Finance in Finland decided to adopt the RL based system *Aurora AI*. This model lets people or businesses create a virtual information avatar of themselves, and triggered by information about life events like university graduation or getting divorced, it suggests service provisions by different kinds of public authorities. These services could support citizens or companies in job searches and other important actions (Valtiovarainministeriö - Finansministeriet, 2019; Kuziemski & Misuraca, 2020).

3.2 Disadvantages of Using AI in the Public Sector

After having focused on the advantages, it is time to consider disadvantages as well. Literature in the fields of social science, law, medical health and public administration extensively describes potential risks of using AI for governance decisions. In this section, arbitrary categories present insights from all of these fields, scoped for downsides which potentially affect citizens and not only the organisation which uses the AISs. This preliminary literature review helps to find what knowledge about these downsides still needs to be developed for meticulous public sector AISs. Based on that finding, a directive for the development of a theoretical construct is set in the last section 3.4 of this chapter.

3.2.1 Biased decision-making

AISs often need to be trained with large data sets. These data sets possibly contain historical biases (Kroll et al., 2017; Floridi et al., 2018; Corbett-Davies & Goel, 2018). Furthermore, AISs can reinforce existing societal biases and consequentially societal polarisation (O'Neill, 2016). The selection of training data itself is a potential source of selection bias (Barocas & Selbst, 2016; Boyd & Crawford, 2012; Corbett-Davies & Goel, 2018). If input data contain biases and errors, or data from different sources are rashly combined, decisions informed by the data analysis could be negatively affected (Janssen, van der Voort, & Wahyudi, 2017). Introduction of AI in public decision-making leads to new biases as well. Classifications by the system might be valid for the population, but not on the individual level (Janssen & Kuk, 2016). If AISs serve for surveillance purposes, runaway feedback loops lurk. Systems steer towards over-surveillance in certain areas or societal subgroups, because data with higher numbers of charges in

these groups feed into the system over and over again (Dobbe, Dean, Gilbert, & Kohli, 2018).

3.2.2 Discrimination

If data sets contain sensitive citizen attributes like gender or racial information, and the AIS uses this information for either classification or predictive purposes, AI possibly fuels discrimination. Individual cases with equal scores on relevant features (e.g. fraud penalties in the past) should have the same AIS risk score for committing fraud in the future. However, due to membership of some sensitive attribute-defined subgroup, one of the individual cases ends up with a higher risk score (Corbett-Davies & Goel, 2018). The same holds true if proxy variables for discriminatory attributes are used, for example ZIP code information (Kroll et al., 2017; Barocas & Selbst, 2016; Whittaker et al., 2018).

3.2.3 Privacy

Combination of different data sources can undo anonymity of each of the separate data sources, scrutinising the privacy of citizens (Boyd & Crawford, 2012). Furthermore, criminals can get access to citizen information through hacking if AISs process sensitive data (Kitchin, 2014). New processing methods of citizen data by AI continuously shed new light on the role of privacy and trust in automated decision-making as well (Janssen & Kuk, 2016). AI's data requirements in general can have harmful effects on citizen privacy (Dobbe & Raji, 2019).

3.2.4 Corruption of the system

AISs are prone to manipulation by users (Janssen & Kuk, 2016; Dobbe & Raji, 2019). If only certain groups of citizens know how to do so, these citizens gain unjustifiable advantages compared to others. Floridi et al. (2018) call this an unequal distribution of AI costs and benefits over citizens. Kroll et al. (2017) provide an example of how an AIS for the assignment of first-choice schools in the United States lead to strategic behaviour in the fill out of preference forms. Parents who knew how to make use of this glitch in the system, gained strategic advantage over parents who did not. In this way, many children with unknowing parents lost out on access of their top-3 school choice.

3.2.5 Malfunctioning of the system

Another disadvantage for citizens is labelling based on patterns that do not exist. Because AISs process immense volumes of data, classifications of citizens are sometimes based on variable relationships without substantial grounding. Mathematical and data-centered systems should not be mistaken for impartial or unambiguous systems (Boyd & Crawford, 2012; Barocas & Selbst, 2016). Malfunctioning or crashing of an AIS which fulfills a crucial societal function can put citizens in danger, for example in healthcare contexts (Floridi et al., 2018). Unacceptable risks also occur in the exploration phase of AI, when a system is first released in public spheres. A well-known example is the fatal crash of a self-driving beta-test model Tesla in Florida (Dobbe & Raji, 2019).

3.2.6 Extensive control

Floridi et al. (2018) raise doubts about the influence of AI deployment on the self-determination of citizens. Nudging of human behaviour by automated systems should

not exceed acceptable levels. Furthermore, there is the risk of harmful exploitation over citizens by the deploying actor of the AIS (Dobbe & Raji, 2019).

3.2.7 Lack of transparency and accountability

Opacity of AISs inflames previously mentioned problems concerning fairness and human autonomy. Another disadvantage for citizens that arises from it, is the lack of governmental accountability (Ziewitz, 2016). Due to the opacity of AISs, citizens generally cannot check whether all data sources that are exploited for the ADM are justified sources for that purpose (Kitchin, 2014). Classification or prediction can be unlawful but leaving citizens with insufficient means to control the governmental institution (Barocas & Selbst, 2016). Informed consent, an important precondition for AIS deployment by governmental institutions, sometimes only exists due to lack of transparency of the system rather than actual consent. Furthermore, implementation of AI can lead to unanticipated results in its context (Dobbe & Raji, 2019). This again scrutinizes a reasonable interpretation of informed consent and ex post accountability (Janssen & Kuk, 2016). Liability and accountability in cases of wrong AI decisions in the public sector are unclear most of the time. To be more concrete: what if a pedestrian gets hit by a public transport autonomous vehicle and it remains unclear who is accountable for this AIS (Wirtz et al., 2019)?

3.3 Possibilities to Mitigate the Disadvantages

The number of possibilities for mitigation or prevention of AI challenges for citizens found in literature is large. This categorised overview of strategies contributes to both the understanding of interviewee answers during the case study and a typology for different kind of mitigation strategies for the governance recommendations.

3.3.1 Technological measures

Corbett-Davies et al. (2018) describe how fairness metrics can be adopted by programmers in order to protect citizens from system biases. The distribution of erroneous decisions over different subgroups of citizens can be such a fairness metric. A calculation metric of the system's predictive power for every citizen subgroup is another example, as is the distribution of risk estimate by the system over citizen subgroups, which is called calibration. However, these authors also warn for statistical limitations of fairness metrics. In fact, designing AI and algorithms to satisfy fairness definitions could even violate minority and majority groups' well-being. Whittaker et al. (2018, p.27) call this overconfidence in fairness metrics a "dangerous sense of false security".

ADM systems in the public sector are often barely or not at all understandable for end users of the systems. For this reason, scholars now dive into the possibility of Explainable AI (XAI). The idea of XAI is creating algorithmic methods which allow for understanding of human experts who are involved in the decision-making. It therefore is a step towards explanation of algorithmic decisions towards citizens, but at the same time, XAI is not equal to a transparent system. After all, the wickedness of decision-making contexts and the complexity of the data-driven system sometimes prevent transparency, even if the AI is explainable to experts (Janssen et al., 2020).

3.3.2 Data measures

A more stringent regime for data processing could prevent discrimination by AISs. If AISs are not trained or fed with so-called sensitive attributes, decisions cannot be made based on non-relevant attributes. This is an anti-classification strategy (Corbett-Davies & Goel, 2018). However, biases often emerge when other attributes appear to be proxies for sensitive attributes (Kroll et al., 2017; Barocas & Selbst, 2016; Whittaker et al., 2018). These proxy variables are harder to filter out of the system's data. The European Commission's High-Level Expert Group on Artificial Intelligence (2020) advise to minimise personal data use in general. A last strategy is to not exaggerate data fusion methods, as the results might diverge from reality too much (Van Der Voort et al., 2019). Up till now it seems to remain unknown where effective data science ends and data manipulation starts.

3.3.3 Monitoring and evaluation measures

Most biases will only emerge after the implementation of AI in its specific context. Therefore, an ex post study of feedback mechanisms between the AI system and its environment is recommended. Furthermore, domain experts in high-stake domains like the judicial system should keep track of the fairness of decisions made by AISs during its functioning (Dobbe et al., 2018). This human review could take the form of a human-machine partnership (Danaher, 2016, p.266). Fallback plans contribute to the robustness of ADM. Especially low-confidence score predictions need the correct remedies (AI HLEG, 2020). At the same time, monitoring impacts of AISs should not put undue burden on the departments or governments which conduct such monitoring measures (Morley et al., 2021). Third parties must have possibilities to audit the AISs as well. Documentation and logging in this context aids reliability and reproducibility. Auditing itself should be based on accuracy, system vulnerabilities, ethical concerns, and accountability (AI HLEG, 2020). An external audit could also offer methodologies to find potential biases in the data sets used (Chowdhury & Sloane, 2020).

3.3.4 Legal measures

Changes in protection for citizens due to changes public sector AI adoption bring "requires additional legislation or a reconfiguration of the legal framework to compensate for the loss" (Hildebrandt, 2015). Some scholars mention possibilities for legal instruments to deal with the drawbacks or uncertainties public AISs bring. Prins (2020) recommends to include legal experts with teams that evaluate and assess public AISs, besides for example ethical and sociological experts. Others recommend to legally oblige governments to strive for public disclosure of their source codes of every high-impact algorithmic decision-making or decision-aiding system (van Eck et al., 2018). This however raises questions about what makes an algorithmic-system to be high-impact or low-impact. Citron & Pasquale (2014, p.33) call for a new aim of legal systems. According to them, it should aim to provide oversight of what algorithmic scoring systems exist in societies, as these can "narrow people's life opportunities in arbitrary and discriminatory ways". Lastly, a legal instrument could be to change laws which protect some data categories and trade secrets, so audits and in-depth assessments of correct functioning of AISs, which should not cause harms to citizens, become viable (Morley et al., 2021).

3.3.5 Organisational measures

Another kind of measures found in literature relates to governmental organisational roles and structures to deal with the downsides of AISs. Governments can set up some sort of internal auditing system with the help of a dedicated team of governmental employees outside the development team. This internal team should gain full access to all relevant material, e.g. the input data (Morley et al., 2021). A more abstract organisational measure is to actively engage with assessments lists and other methods aimed to safeguard critical reflection on meticulous use of AISs. In that way, an organisational culture arises which is more committed to trustworthy use of AISs (AI HLEG, 2020). Other scholars discuss the internal appointment of specific employee roles to carry responsibility for the handling of different categories of risks public sector AI carries, for example privacy risks. Such roles could for that example then be data controllers (Hildebrandt, 2015) or Data Protection Officers (AI HLEG, 2020).

3.3.6 User engagement and citizen agency

When citizens themselves encounter safety issues for AISs, flagging methods for reporting of these issues must be in place. The system must provide citizens with options to opt out and requests to delete their personal data as well (AI HLEG, 2020). Transparency must contribute to citizen understanding of AISs their governments use. Transparency takes many forms, for example public disclosure, value-centered design or educational efforts (Ziewitz, 2016). When citizens are informed about ADM, they should always be notified about the model's contributions to the decision (AI HLEG, 2020). The same holds for communication about the potential and limitations of the model to model users or citizens (Kolkman et al., 2016). Some also point towards the need for a possibility for citizens to address their specific questions about the algorithmic contribution to decisions (Algemene Rekenkamer, 2021). Citizen agency could also take the form of direct participation in AI development practices, which consequently leads to higher acceptance of public sector use of ADM systems as well (Kolkman et al., 2016).

3.4 Chapter Conclusion

This chapter presents the literature review results which belong to research phase 1 and are necessary to answer two sub-questions, the first one being:

What are the advantages and disadvantages of using AI models in the public sector and what types of mitigation measures are available to deal with the disadvantages?

Public sector use of AI brings advantages which are inherent benefits for the citizens that governments must serve, as well as benefits to governments as an organisation, with limited resources to fulfill their task just like any other organisation[1]. Inherent citizen benefits are the benefits of ADM and AI technologies for citizens relative to conventional decision-making by governments. Telling examples of inherent benefits for citizens are decreased waiting time for governmental decisions and the potential of more objective decisions in some cases. Section 3.1.2 presents a complete overview. Advantages for the governmental organisations itself are positive influences on the efficiency and effectiveness of their decision-making. Section 3.1.1 elaborates on both of these types of advantages. Understanding the advantages helps to prepare for the case study interviews and later to understand the governance requirements and trade-offs the CoA have to make.

[1] I can hear you think "What about Amazon?!" Well, they are the exception that proves the rule.

The disadvantages found in the literature up til now belong to the following categories: biased decision-making, discrimination, privacy issues, corrupt system functioning, system malfunctioning, extensive control over citizens by governments and lack of transparency and accountability. These categories from section 3.2 are no exhaustive representation of downsides, but serve to develop an understanding of what can go wrong and what knowledge is still lacking. Governments also have possibilities to deal with AIS downsides. Section 3.3 presents a typology of such mitigation measures found in the literature: technological measures, data measures, monitoring and evaluation measures, legal measures, organisational measures, and user engagement and citizen agency. This typology can later be used to interpret the mitigation strategies that the CoA uses in its daily practice.

The second sub-question of this research phase relates to the lacking knowledge after having studied the disadvantages of public sector AI use:

Which knowledge of these disadvantages is lacking and what kind of theoretical construct would contribute to this knowledge?

The first aspect that appears to be missing, is structuring the information about disadvantages. The arbitrary categories are now strongly varying and to base governance measures on this preliminary overview would be impracticable. If a TU Delft supervisor strikingly says that a bookshelf seems to be overthrown, it is time to take action. The question then arises what kind of structuring would be helpful. To answer that question, besides critically reflecting on this chapter's overview, preliminary discussions with people with different backgrounds and expertise were held. These expertise include legal, AI development and governance expertise. Appendix A contains table A.1 which presents information about such discussions, as well as some additional relevant events that were visited.

In this first overview, different levels of AIS downsides are visible, they are not all technologically originated. To understand how public sector AISs potentially harm citizens, it is not enough to come up with generic downsides. Downsides do not stand on its own, but are context-specific and actor-specific. The theory in the next chapter must address how the actors from relevant contexts play a role in how downsides emerge. Furthermore, the downsides do not all come to light at the same time during the development of AISs for governments. There is a need to understand what are relevant demarcated time steps in the development practices and in which time step the downsides start to play a role as well. Knowledge about *how* the downsides actually emerge still lacks as well. If the theoretical construct considers the relevant contexts, it becomes more clear how the downsides arise in these contexts. Lastly, it remained unclear in this chapter how the downsides of public sector AISs can result in actual harms for citizens. This must be illustrated to make clear how the downsides may eventually lead to actual citizen harms.

Chapter 4

A Conceptual Model Representing AI System Vulnerabilities

This chapter reports about research stage 2: Building a relevant theoretical construct to understand the origin of public sector AI vulnerabilities. Chapter 3 presented a general overview of merits, as well as disadvantages for citizens caused by adopting AI technologies in the public sector. But in order to better understand how public organisations can reap the benefits of AI technologies without causing citizen harms, meticulousness of AISs is necessary. From conducting the literature view, it becomes apparent that the ultimate harms for citizens not only originate in the technological model itself. And to build meticulous AISs in the public sector is to not only understand what general weaknesses of ADM systems exist, but also be able to prevent or mitigate risks by establishing governance measures in the right place at the right time. The key elements for this chapter's theoretical framework to address based on the knowledge gap determined in the previous chapter are:

- Understanding how the potential harms for citizens emerge
- Context-dependent and actor-dependent interpretation of downsides of public sector AISs
- A well-structured construct, rather than just one-by-one presenting strongly varying downsides

This chapter contributes to the goal of addressing these knowledge gaps by presenting a conceptual and layered 'onion' model of AIS **vulnerabilities**. Vulnerabilities are not the same as the downsides from chapter 3. Downsides are merely generic challenges encountered when governments use AISs. Vulnerabilities are context-dependent elaborations of such downsides. The vulnerabilities encompass many different potential shortcomings of the system which may ultimately result in harms to citizens. If these vulnerabilities actually arise in public sector AISs, this eventually can lead to harms for citizens. This is illustrated at the end of every model layer.

4.1 Conceptualising Public Sector AI Vulnerabilities from a System Perspective

The conceptual interpretation of AI vulnerabilities, which this model essentially is, is by no means a complete coverage of all potential vulnerabilities of public sector AI. Rather it is a tool which aids to better understand what the root causes of potential citizen harms are when public organisations use AI technologies in their operations. These root causes are made context-dependent. By doing so, the model is a good preparation for the case study and a base to answer the main research question.

Figure 4.1 presents the conceptual onion model for this chapter[1]. Every layer represents a contextual layer of the AIS from which vulnerabilities possibly originate: the context of the model itself, the model deployment context, the political-administrative context and the societal context. The idea to distinguish between these four contexts appeared during the preliminary talks with experts - see chapter 3 and appendix A. These talks also served to iteratively sharpen the ideas and literature search for this conceptual framework and the set of presented vulnerabilities.

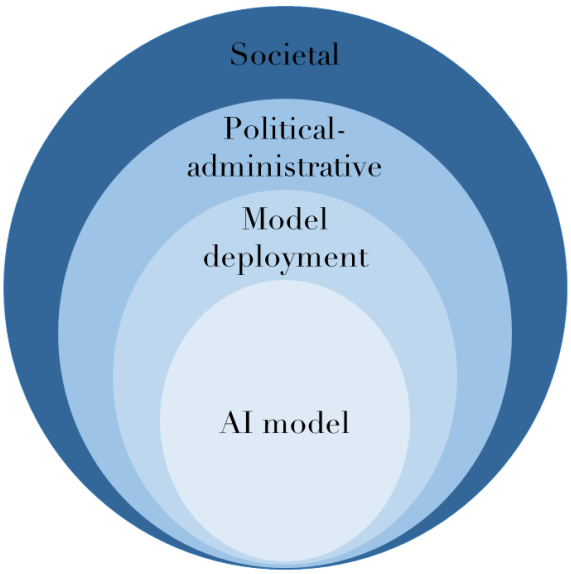

Figure 4.1: A conceptual model of public AI systems vulnerabilities

4.1.1 The Inherent Logic and Layers of the Onion Model

The conceptual model has an inherent logic which makes it a usable construct to interpret AIS vulnerabilities. First of all, the contexts and actors in the contexts themselves. If a government decides to buy or develop some AIS, the first source of potential vulnerabilities is the model itself, and the developers involved during the modeling. There is a department which deploys the model for its daily operations, where other problems may occur. Then there are the political-administrative overarching policy objectives and general needs which are involved with everything the specific government does, for example legal demands for governmental services and the city council striving for equality within the city. Lastly, the societal context represents the actors who are ultimately affected by the AIS, namely citizens, where other vulnerabilities appear. The

[1]Indeed, it is actually not an onion. Luckily it does have layers.

contextual layers together constitute a unity which covers the scope of the AIS. They are not separated by firm limits but flow into each other, as the actors from different contexts are connected in real-world as well.

Relationships between the contextual layers from outside to inside are hierarchical: this means 'should be able to have power over'. For example, the domain experts and end users should be able to determine what specifications the AI model needs to have. Reflective, the other way around the relationships carry the meaning: 'should contribute to reaching the goals of'. E.g., the political-administrative actors in a public organisation in essence have the democratic mandate to serve the goals of civilians and other societal actors. From this inside-outside perspective, it also interesting that every layer is further drifted away from the model itself. This shows how in real life the actors from these layers are further away from access to the AI model as well. The societal context is the outer and upper layer, which reflects the fact that to serve the goals of society should be the ultimate purpose of the AIS.

4.1.2 Structuring the Vulnerabilities Within the Four Contexts

During the iterative process of discussing vulnerabilities with experts and reviewing literature, it appeared that there is a need to understand during which stage of AIS development the vulnerabilities originate. This helps to further understand the vulnerabilities, but is also useful to establish reasonable governance requirements later in the process of this thesis. Two AI model development phases form the basis of the categorisation of vulnerabilities. These are *Before implementation* and *After implementation*. This distinction is based on the Hard Choices in AI framework, a framework intended "to deliberate critically and constructively about normative indeterminacy." (Dobbe et al., 2021, p.22). The four phases presented in that framework are:

Problematisation - This phase concerns the understanding of a situation which might benefit from an AI solution. It is necessary to problematise, i.e. determine the different interests which are relevant for the situation. Furthermore it must become clear which municipal departments the AI solution serves and which department is responsible for administration.
Featurisation - AI models are abstractions of the public sector realities they are used for. In the featurisation phase of development, involved actors determine what the model needs as input and what outputs are required to come up with adequate decisions.
Optimisation - This phase concerns finding parameters and features of the model which constitute acceptable input-output behaviour. Data sets are necessary to train the model, as well as evaluation criteria. Both model developers and model deploying actors seek to find out whether the model functions as expected and serves its goal.
Integration - The model gets implemented in its deploying department. Unexpected or unwanted mechanisms may occur through interaction of the model with its domain of application. The governmental agents involved with the AIS should think of the forms of agency affected citizens and other actors have if the AIS unduly disadvantages them and feedback mechanisms for affected people.

The four presented phases describe relevant considerations in every model development stage. The most interesting distinction for vulnerabilities arising is between the phase before implementation of the model, and after implementation of the model. That is because the governance requirements mainly concern (i) choices to be made about why a model is necessary and how it should be modeled, and later (ii) how to deal with the integration of the model into the governmental operations. Hard Choices' considerations are thus grouped into two categories. Another advantage of bringing back the number of development phases categories, is that the categorisation becomes more intuitively understandable. The two development process categories are:

- **Before implementation**: vulnerabilities which arise during the *problematisation*, *featurisation* and *optimisation* of AI models
- **After implementation**: vulnerabilities which arise during the *integration* of AI models

The coming subsections present the vulnerabilities in every relevant context of the AIS, represented by the four layers in the conceptual framework of AISs: AI model context, model deployment context, political-administrative context, and societal context. Every model context is divided into two panels, based on the categorisation as just presented. Figure 4.2 presents the categorisation of vulnerabilities in each layer of the conceptual onion model. Each layer description comes with an illustration to clarify how vulnerabilities may results in actual harms to citizens. If deemed necessary for clarity, vulnerabilities are also illustrated themselves.

	Before implem.	After implem.
AI model	I	II
Model deployment	III	IV
Political-administrative	V	VI
Societal	VII	VIII

Figure 4.2: Vulnerabilities within every context: before and after implementation

4.2 AI Model Vulnerabilities

This section presents the vulnerabilities that emerge from the AI model itself, as well as from the actors involved with developing the model. An illustration of how vulnerabilities from the AI model context lead to actual harms concludes the section.

4.2.1 Panel I: AI model, Before implementation

Overfitting and underfitting
Overfitting is a commonly encountered problem in building AI models. The vulnerability lies within the parameter set for the AI model. This bounded set of data does not represent the whole population. If the model is made too complex, i.e. too many features are used for the model's purpose, overfitting occurs. The large number of added features start to explain patterns and effects that are not telling for the real world situation, but caused by the so-called 'noise' in training data (Kroll et al., 2017). During the training phase of your modelling trajectory, the error residual gets lower.

However, if the model is tested in real-world situations, the test errors increase, on the contrary. The AI model predicts or classifies in a way that does not do justice to the real problem (Dobbe, 2020b). Underfitting is another encountered problem for AI models. An underfit model is an oversimplified model which uses too few input variables and parameters to explain the real-world phenomenon based on the training data set (Ghahramani, 2015).

Poor data quality
For public organisations to build AI models, large amounts of data are necessary. The quality of these data sets varies: data can be incomplete, inaccurate, outdated, biased or untruthful. Low quality data or poor quality data poses major challenges for governmental bodies which seek to use AI and probably lead to failures of the models (Wirtz et al., 2019). It is for the reason of this vulnerability that the Netherlands Court of Audit marks data quality and data set maintenance as key audit subjects for public sector AI (Algemene Rekenkamer, 2021). If AI developers want to use their model for predictions, the used data must meet certain conditions to be used for inference. Not meeting these conditions might result in faulty decisions and hence harms to citizens. A vulnerability of public sector AISs therefore is failing to meet the conditions that belong to the goal and architecture of the model (Dobbe, 2020b).

Training data gaps
Narayanan (2019) composes a crude categorisation of type of models based on their classifying or predictive performance. For many models which seek to predict social outcomes like criminal recidivism or at-risk kids, it is very hard for AI models to reach acceptable performance standards. This is despite the fact that there is much available training data. Sometimes, there is no training data available which would aid the AI model in a sufficient way. Training data can simply not contain all relevant information and is sometimes restricted for modelers. Some real-world situations happen so infrequently that it becomes too complex to compensate for by creating synthetic data. Questions must then be asked about the appropriateness of AI or ML models for these uses (Dobbe & Raji, 2019).

Mathematical impossibility of combining observational fairness strategies
Bias in predictions or classifications is an often mentioned problem for ADM. Although there is no single method to prevent or mitigate bias, there are ways to mathematically express bias in outcomes and using these, set goals in the algorithms to minimise bias. Corbett-Davies & Goel (2018) elaborate on three of such mathematical ways to express bias, which they call observational fairness strategies. However, it is impossible to optimise for multiple observational strategies at the same time. For example, one fairness strategy is to not use sensitive attributes as gender and race. Another observational strategy is classification parity, which is to equal predictive performances across groups with the same sensitive attributes. Intuitively, these observational fairness strategies cannot be optimised simultaneously.

Computational systems are biased towards quantifiable factors
The nature of computational systems like AI obviously requires the input for decision-making to be quantified. Consequentially, the AI model is biased towards system features that lend itself to quantification. Some 'soft' factors that might be important for nuanced decisions are then left out of the decision-making process (AI Now Institute, 2018).

Sensitive and personal information in available data
Data sets possibly contain sensitive information which should not be available to modelers, legally or morally, or lead to discriminatory outcomes. Modelers who want to avoid discrimination by their model can leave out sensitive attributes like nationality or race. But when other data entries highly correlate with sensitive information, and

these so-called proxies for sensitive attributes are not left out from the model, AI models can still be discriminating towards certain population subgroups. For example, zip codes and language handling aid to predict a person's ability to pay back a loan, but correlate with sensitive attributes like ethnicity (O'Neill, 2016).

Biased data collection
The process of data collection might be biased as well. A clear definition found in literature is "Biases introduced due to the selection of data sources, or by the way in which data from these sources are acquired and prepared" (Olteanu, Castillo, Diaz, & Kıcıman, 2019, p.13) This leads to an AI model which is vulnerable for decision-making based on biased information. Civil servants sometimes manipulate by not collecting crucial information because this information might give them a bad look. For example, police officers in New York persuaded victims not to report assaults, so that the crime numbers in the city looked better than they were (Richardson, Schultz, & Crawford, 2019). Or some members of populations are more likely to respond in data collections - this is called *response bias* - and measurement of relevant information influences the data itself: this is called *measurement bias* (Dobbe, 2020a). A last type is *selection bias*: the bias that occurs because choosing certain data sets for modeling influences the observations and therefore modeling results during development (Olteanu et al., 2019).

Acceptability of model implementation
Another dilemma and thus vulnerability for governments is to find the right moment, performance standard or accuracy of their application to go from lab phases to real-world implementation. The model is never finished and errors or complaints could lead to new insights for further improvement of the model. An example from New York is a DNA matching AI model that was expected to perform according to standards using a sample of 20 picograms. Later in time it appeared that the sample size needed to almost double before acceptable performance of the model was reached (Kirchner, 2017).

4.2.2 Panel II: AI model, After implementation

Mathematical accuracy of decisions
Another concern for AI models is accuracy of decisions, seen from a technological perspective. Accuracy then refers to the rate of decisions from the model that is correct (Dobbe, 2020a). There are several metrics for accuracy, for example the False Positive Rate or the False Negative Rate. Metrics which indicate how much of the variance is explained by the model, also tell us something about the accuracy of the AI model. An example of such a variance metric is the R2 statistic of a model: which proportion of the total variance is explained (Dobbe, 2020b). Some scholars criticise AI models for their low accuracy of decisions. E.g. Narayanan (2019) calls many AI models which seek to predict social outcomes "snake oil", because they do not outperform simple manual scoring methods which use a very limited set of features.

Different accuracy of decisions for different subgroups population
A vulnerability which concerns both accuracy and bias is the distribution of errors over different population subgroups. Research shows that AI models sometimes have higher accuracy of outcomes and decisions for certain subpopulations. Accuracy scores of predictions for groups with different ages, sexes or socioeconomic backgrounds can vary and if they do so in a structural way, the algorithms are biased (Dobbe, 2020a). An example is the disparity of error distribution from facial recognition systems comparing between racial and gender features (Raghavan, Barocas, Kleinberg, & Levy, 2019). This leaves the vulnerability of biased or discriminating ADM.

Problems with validation
An AI model needs sufficient validation in order to be used by a public organisation in a responsible way. If the model validation is of insufficient quality, outdated or lacking completely, the AI model might not perform in the way it is intended to do. Unwanted consequences of the model can result in citizen risks. The importance of validated working from a model holds for the entire lifecycle of an algorithmic system, but especially after deployment or implementation of the system (AI HLEG, 2019).

Inadequate fail safe mechanisms and plan B procedures
Like all technologies, AI models sometimes fail in their functioning. If you know the system makes mistakes, you have to anticipate them and organize fail-safe mechanisms that are activated when things go wrong to safeguard the system and affected stakeholders. The development team therefore must set up plan B procedures as a back-up for the system's functioning. If the development team fails to do so, citizen harms may occur. A severe example of failing plan B procedures is the lack of safety fallback plans for self-driving cars that were still in beta-testing, causing deadly accidents in the United States (Dobbe & Raji, 2019).

4.2.3 Illustration of harms due to AI model vulnerabilities

A telling example of technical bias of AI models is found in Williamsburg, New York. A man called Mayer Herskovic was accused of beating up a black man, fitting into the picture of the ongoing struggles between the black communities and Hasidic Jewish communities in Williamsburg. A mixture of the battered man's DNA and Herskovic's DNA was found on the shoe of the victim. However, Herskovic's lawyer made a successful appeal. Because of the fact that Hasidic Jews in Williamsburg live so isolated and share many of the same ancestors, the accuracy of DNA testing for Hasidic Jews is much lower (Kirchner, 2017). This is a clear example of a vulnerability in the form of technological bias of an AI model which (almost) leads to unjust harm to a citizen. The AI Now Institute (2019) concludes that the Chief Medical Examiner of the city used its "faulty algorithm" for thousands of criminal cases.

In Idaho, people with disabilities saw their social benefit payments drop drastically after the state adopted a new ADM system to calculate these benefits. This lead to "horrific living conditions for participants who were no longer receiving enough hours of in-home care and services" (Richardson, Schultz, & Southerland, 2019, p.28). After a court case, the Medicaid program had to improve the algorithms so that people received enough funds, as well as put in place extra procedures for fallback and compensation of already impacted individuals. This example illustrates the potential vulnerabilities of model developers who do not have enough domain knowledge to understand the impact of their creations, as well as lacking plan B procedures.

Another case from the United States illustrates the vulnerability of problematic validation of models. A so-called Structured Assessment of Violence and Risk in Youth (SAVRY) reports about the risk of violence from young defendants. This automated risk profiling can make the difference between being incarcerated or being given probation. A young defendant pleaded guilty for a robbery in order to receive probation. However, due to the SAVRY assessment of his profile, he had to go to jail. His lawyers later found out that the performance of the SAVRY assessment had not been validated in a sufficient way, making these high-impact assessment very debatable. One of the validations was more than two decades old, the other was an unpublished master's thesis (Richardson, Schultz, & Southerland, 2019). Although the latter can be of extraordinary quality too, of course.

4.3 Model Deployment Vulnerabilities

This section concerns the AIS vulnerabilities arising in the governmental model deployment context, again categorised for their moment of arising: before or after implementing the model, panel III and panel IV from figure 4.2 respectively. The illustration at the end of the chapter shows how model deployment vulnerabilities have led to harms for citizens.

4.3.1 Panel III: Model deployment, Before implementation

Incorrect input-output relationship
AI-based predictions or classifications can be based on patterns or correlations that may exist in the data, but do not do justice to reality. This threatens the prerequisite that public decisions about citizens are based on the matching information and therefore justifiable. An example can be found in job application procedures and predicting criminal recidivism. Raghavan (2019) scrutinizes models which use facial recognition to classify emotional states of job applicants. This facial expression-emotion relationship is highly unreliable, especially when considered from a cross-cultural perspective. The same happens when pre-arrest data serves as input to a recidivism prediction model in the United States: arrests do not exclusively come from criminal activity, but also from biased police practices and policy activity in certain areas (Narayanan, 2019; Dobbe, 2020a).

Removing humans from decision-making can be a compromise on quality
If ADM systems replace the existing practice of human agents being in contact with citizens in need of service and care, the quality of service provision can decrease rather than increase. If traditional benefit allocations are based on one-on-one assessment by human agents, the assessment is not only based on numbers, but on meaningful dialogue as well. If this dialogue with citizens is replaced by AISs, the public organisation potentially eliminates valuable input to the decision (AI Now Institute, 2018).

4.3.2 Panel IV: Model deployment, After implementation

Models unequipped for changes in application domain reality
Another vulnerability arises when an AI model in the public context, with established data inputs and decision rules, is not able to deal with unforeseen changes and variety in the application reality. In fact, AI models are often criticised for their susceptibility to do so (Janssen et al., 2020). AI developers must pay great attention to the adaptability of their systems. During a tech demonstration in the CoA, an AI practitioner questioned the reliability of his image recognition system during periods of snow. Natural changes like these resound in the input features of the model and therefore can change the model behaviour in an unwanted way (Dobbe & Raji, 2019).

Emergent bias
Emergent bias is a form of bias which arises in the application environment of an AIS. AISs in public operations can show emergent behaviour, i.e. effects of using the system that were either unexpected or not explicitly predetermined by the organisation which uses the system (de Bruijn & Herder, 2009). In other words, the AIS behaviour in the real-world application domain causes unanticipated effects. Shifts in the application reality or context of use lead to difficulties for specific groups of citizens affected by the model. According to Friedman & Nissenbaum (1996), it is a form of bias which arises due to changes in for example societal knowledge or population after the model has been implemented. Dobbe et al. (2018) point to feedback loops between the model environment and the model itself as important causes of emergent bias.

Lack of model understanding with deploying department's workers
AI models are often intended to support human decision-makers in their processes. But if the models completely lack transparency or are too hard to understand for the human agent, the outcomes of this ADM can be too hard to interpret for the human agent (Janssen et al., 2020). If the support systems are not understood correctly, quality of decisions might decrease because the quality of decisions cannot be checked for within the organisation. Van Eck et al. (2018) call this the 'first line of control' within the own hierarchy. This quality check is necessary because public AI often concerns societal issues which various actors can have different opinions on and AI models will not find perfect solutions (Narayanan, 2019).

Negative impact on workforce
Citizen harms can also take the form of job insecurity for public sector employees. If human agents, working for the department of application of the new AIS, are replaced, we find a new form of citizen harm. This idea matches the broader societal discussion and anxiety about the replacement of human workforce by AI (Wirtz et al., 2019). Another negative impact can arise when many workers are needed for badly paid and mundane, boring tasks of for example labeling AI training data or reviewing automated decisions manually (Dobbe & Raji, 2019). Professionals could encounter significant infringement of their professional freedom of action, especially in large-scale public organisations which e.g. decide about benefit receivings and fines (van Eck et al., 2018).

New form of citizen dependence on unpredictable public sector interpretations
Although AISs are often applauded for claims of objectivity, they possibly also create a new form of citizen dependency on human vagaries. If some professionals put their faith in algorithmic decision support and others don't, a new unpredictability of public sector operations is created. It is likely that this trust in the AIS depends on the professional's knowledge of the system and generic attitude towards technology. Richardson et al. (2019) describe a court case in which a judge did not accept a questionable AI-model for risk profiling as evidence for the district attorney. However, this ruling was not used as jurisprudence by other juries. Whilst some citizens had to face evidence from this SAVRY risk assessment system, introduced in section 4.2.3, others did not.

4.3.3 Illustration of harms due to model deployment vulnerabilities

Frictions between AI developers and domain experts are illustrated by, again, a case with DNA matching software from Kirchner (2017). Eli Shapiro was a technical leader of a DNA lab. He found himself under pressure of accepting new AI technologies for DNA testing. This DNA expert decided to retire early from his job as technical lab leader after the introduction of a new AI technology. That is partly because he was stressed about signing off automatically generated reports about DNA from which almost no one in the lab understood the working. This anecdote illustrates the lack of trust and understanding which human experts can have with regards to AISs in public sector contexts. The domain expert expects the AIS to make inaccurate decisions which potentially bring critical harm to citizens in courtrooms. It also illustrates the potential negative impact on workforce as a vulnerability, because Shapiro was under pressure and decided to retire early.

In California, navigation apps like Waze were not equipped for the changing application context of road blocks due to wildfires. While large parts of the State were on fire, the sudden danger of fire lead to an unanticipated problem. Parts of towns that were abandoned due to the imminent flames, logically encountered less traffic. But

navigation apps are trained to advice people to hit roads that are empty in order to decrease travel time. This led to multiple citizens being sent to roads that were already blocked by the State because of the wildfires (Mak, 2017). Due to the inadaptability of the systems, citizens were put in danger. This reflects the citizen harm which can be the result of the vulnerability of training AISs that might be efficient normally, but are not as adaptive as human agents.

4.4 Political and Administrative Vulnerabilities

The third model layer concerns the vulnerabilities emerging in the political-administrative context of governments. This includes both the political actors like a city council and administrative actors like municipal management teams. Again, the split for vulnerabilities emerging before and after model implementation is made - panel V and VI from figure 4.2 respectively. An illustration of citizen harms caused by political-administrative vulnerabilities finishes the section.

4.4.1 Panel V: Political-administrative, Before implementation

Presumption of innocence compromised
Amongst other scholars, Van Eck et al. (2018) warn for the potential infringement of the fundamental law principle of presumption of innocence by self-learning computer systems in the public context. Hildebrandt (2015, p.8) elaborates on this issue and even warns for the AI-fueled reversal of the principle, which she calls the "automation of suspicion". The presumption of innocence must ensure that a person is believed to be innocent as long as there is no strong evidence which suggests the opposite. Automated flagging of citizen behaviour based on data patterns in law enforcement and intelligence can lead to a presumption of guilt, due to the inherent need of technology to flag based on input data. This is a lost legal protection shield for citizens. The innocence presumption is related to AI privacy risks as well, since the right to privacy for citizens should be a first firewall against excessive control and detection technologies by governmental agencies (Hildebrandt, 2015).

Discord with existing legislation and concerns about proportionality
Relative to conventional decision-making, the use of AI to make or support public decisions is still in its infancy. The possibilities caused by high-end data processing algorithms seem boundless. See section 3.1 for a first glimpse of this spectrum. But the large scale data use comes at a cost as well. Sometimes governmental agencies are unaware of the illegitimacy of their digital profiling practices and data use (O'Neill, 2016). This can lead to unlawful and therefore unfair treatment of citizens. An agency from Michigan disallowed food assistance programs based on felony warrants in 2013 for example, thereby violating multiple federal laws and Constitutional due process requirements (Richardson, Schultz, & Southerland, 2019). Furthermore, AI models are criticised by jurists being disproportionately intrusive relative to the advantages they bring. Interference with citizens' personal life is only legitimate if the privacy infringement is legally considered to be inferior to the achieved goal of the model (Huisman, 2020).

Inadequate financial resources allocated to AI development teams
Section 4.2 presents the vulnerability of a discrepant supply of and demand for technological developers in the public sector. This leads to capacity problems in development teams. A related vulnerability found in this political-administrative model layer, is the constrained budget allocation to public AI programs. Governmental budgets are still mostly spent on conventional service provisions (Sousa et al., 2019). If governmental bodies do want to jump on the acsAI bandwagon, but are not prepared

to make the necessary budgetary choices, this vulnerability for underperforming public AI teams arises. Financial feasibility therefore is of upmost importance in the public AI context (Wirtz et al., 2019). This vulnerability might be especially present to resource-constrained governments and public agencies. In the context of farming technologies, Mateescu & Elish (2019) state that new digital technologies require extensive financial resources. Inequalities between resource-rich and resource-poor agents in the field are entrenched by AI possibilities. In this research' context, this could mean that governmental bodies with less financial power could deliver poorer quality automated decisions to their citizens.

Relying on AI models undermines broader political debates
Immediate observations or common sense are not enough to deal with the complexity of modern urbanised areas. AISs aid in this sense making. However, to provide meaning to the outcome of AI models essentially is a political matter, not a technological one. Political-administrative actors must not only use AISs to pursue specific goals in their domain of interest - think of combating traffic jams to increase mobility. They have to also think about the effects of model use on well-being in the city as a whole, for which they are politically accountable (Johnson, 2020). However, to model complete cities just as composites of individual components leads to poor understanding of the functioning of a city as a whole, despite the promise of smart cities to integrate all generated data from the urban area. Using AISs will not overcome the complexity of governance dilemma's in cities. AISs can lead to the downplaying of political debates about what constitutes a city which functions in favor of its citizens, by taking a technological approach of quantitative analysis of an artificially sub-component divided city (Johnson, 2020). This might lead to the vulnerability of taking decisions about public values and policy objectives out of political debates which can be understood and influenced by citizens.

Weak relationship between AI deployment and public policy goals
The Dutch audit institute Algemene Rekenkamer (2021) asks whether AI is becoming a means in itself rather than a tool to solve public issues. If public sector organisations like the CoA change their operational strategy, business rules and maybe even existing legislation to fit the automated systems they use, AI potentially becomes an end in itself rather than a means to an end. Van Eck et al. (2018) warn for this effect. Public algorithm developers, who they call system-level bureaucrats, replace the civil servants responsible for quality of public operations and first-line service quality checks. The first line of defense for citizens which the organisation itself was, has now vanished. Citizens in trouble immediately have to reach out for legal protection. Public management's primary goal in this way is not to rightfully deal with individual citizen cases anymore, but to steer production. Laws and operational rules and standards are adjusted and harmonised to fit the algorithmic systems best.

4.4.2 Panel VI: Political-administrative, After implementation

Diminishing possibilities for corrections of mistaken decisions
The opportunities of AI for public decision-making or informative support are grounded in the use of large and multiple sets of data, see section 1.2.1. Public AI models might be vulnerable for creating harms if the required data sets come from many different administrative teams. Automated decisions from one team depend on information from others, and their decisions subsequently input important automated decisions from the next. In this way a chain of data exchange between multiple teams within the organisation, or multiple public organisations, starts to develop. This chain of information is hard to deal with in terms of risk management. If mistakes are made somewhere in the chain, they are hard or impossible to correct. The important notion of retroactivity can then be compromised (van Eck et al., 2018).

Shifts in discretionary power of public agents
Another institutional vulnerability is related to discretionary power. Before the introduction of self-learning computational systems in the public sector, public professionals were the agents that had the discretionary power to make important decisions about many individual citizens' cases. Think about social benefit decisions or housing decisions. If this power shifts away from civil servants to developers of public IT systems (van Eck et al., 2018; Wieringa, 2020), it is questionable whether the current administrative institutions provide the right checks and balances for this discretionary power for developers. This challenge becomes even more critical when external companies develop the AI models which are then to be implemented by others (Richardson, Schultz, & Southerland, 2019).

Disconnection between AI systems and organisational oversight
ADM and information systems often cross organisational scales in public organisations. That is, they are used, developed, adjusted and controlled by different organisational entities. In some cases, none of the involved entities has a clear and complete overview of what data sets are processed and how the AISs exactly work (Richardson, Schultz, & Southerland, 2019). It is therefore notoriously hard to create internal organisational accountability for risks of these systems. Veale et al. (2018) ask the question whether some ML models might be 'airdropped' from labaratory settings into the political, noisy and complex realities of public sector operations. There is a movement towards centralising data sharing and utilisation for AISs in the public sector. If these same organisations create siloed agencies for oversight capacity, a vulnerability of high-stake decisions with lacking oversight or control for citizen harms emerges (Richardson, Schultz, & Southerland, 2019).

4.4.3 Illustration of harms due to political-administrative vulnerabilities

In December 2020, an allegation of social assistance fraud punished with a 7000 euro fine lead to social unrest in the Netherlands. A woman was accused of fraud because her mother sometimes brought her groceries, and she did not notify her municipality Wijdemeren. The hard-line fraud accusation was especially painful after the Dutch *Toeslagenaffaire* described in section 1.1.1. Wijdemeren started the fraud investigation after begin notified by an authority which was unknown to most citizens, intelligence agency "Inlichtingenbureau" (van der Linde, 2021).

This Inlichtingenbureau works for the united Dutch municipalities *VNG* and the Ministry of Social affairs and serves to detect social assistance fraud. The authority connects many different data sources from public agencies: data about car ownership, academic administration and inheritance receiving for example. The Inlichtingenbureau analyses the data and notifies municipalities, which are the social assistance providers, of suspicious individuals who might be committing fraud (van der Linde, 2021).

This example illustrates different vulnerabilities described in this political-administrative layer. Dutch privacy authority *Autoriteit Persoonsgegevens* casts a serious doubt on the legality of the data processing in light of existing Dutch privacy laws (van der Linde, 2021). The lawyer of the accused fraudster also claimed that the information about the groceries was not obtained in a juridically correct way (Dijkstra, 2021). This corresponds to the notion that AISs are sometimes not in accordance with existing legislation. The presumption of innocence is also compromised if all social assistance receivers are checked based on non-criminal information sets. It is also interesting to consider this illustration of citizen harm from the perspective of AI and public policy objectives. Does having the opportunity to process all these different data sources automatically mean that we as a society now want to tighten up our fraud detection?

Lastly, a warning for the impossibility to correct mistakes in systems like these because public servants lack oversight, is given in (van der Linde, 2021). This notion suits the vulnerability of disconnection between AISs and organisational oversight in this section.

4.5 Societal Vulnerabilities

This last model layer-section shows the vulnerabilities in the societal context of public sector AISs which emerge before and after implementation (panel VII and VIII). The last subsection again is an illustration of citizen harms arising from the coming vulnerabilities.

4.5.1 Panel VII: Societal, Before implementation

Unverifiable and faulty modeling thresholds
In programming, decisions rules are based on thresholds. If logic conditions are met, i.e. a certain value exceeds a threshold or not, the model will or will not execute some next part of the code (Kroll et al., 2017). For example, text classifier models use certainty thresholds to decide whether words can be placed in certain categories or not. A societal vulnerability emerging from this decision logic is that the threshold, which has a decisive impact on the classification or prediction, can be opaque, incorrect or not shared with citizens. This leads to lack of social control on accuracy of the models, as decision thresholds influence the false positive rate and false negative rate (Dobbe, 2020a). Margins of error are used as thresholds, and sometimes later prove to be surprisingly high and based on incorrect information (Kirchner, 2017). An example in Florida shows that a person is convicted of selling crack based on a facial recognition system. Out of 5 suspected people, the model ranked the convicted with a 1 star score of matching the photo taken from the crack seller. Other suspects received a 0 star score. Only in court it became apparent that the system used a 5-star ranking system for possible photo match with unknown internal reliability. This raised serious doubts if it was lawful to use this 1 out of 5 score, despite the fact that others received a 0 star score (Richardson, Schultz, & Southerland, 2019).

AI-inflamed hardening of governmental decisions about citizens
Amongst others, the influence of consciousness and emotion on humans is one of the factors that distinguish human decision-making from ADM. Human decision-makers do not act only instrumentally. Comparing the two sorts of decision-making, it therefore "stands to reason that their value judgments may differ in certain situations" (Wirtz et al., 2019, p.604). To understand emotional expressions or feel emotions might be a very important shortage of AISs in public context decisions, which have moral aspects. If governmental agencies adjust their logics to ADM systems (Richardson, Schultz, & Southerland, 2019), citizens might experience an unwanted hardening of public service provision.

Privacy infringements
AISs often need large amounts of data to be trained, tested, validated, and used. This often concerns personal data from citizens. Intrusion of citizen privacy by data collection and processing methods which are importunate and unwanted is therefore possible. The vulnerability of privacy infringement is especially interesting because of the fact that there are trade-offs between maintaining people's privacy and optimising the AI solutions in terms of smartness (Streitz, Charitos, Kaptein, & Böhlen, 2019). Another famous trade-off for this vulnerability is the trade-off between privacy infringement and increasing safety through algorithmic systems. This is a trade-off which according to Rahwan (2018, p.4) "society must resolve".

Increasing information asymmetry between public agencies and citizens
AISs increasingly contribute to governmental decisions about citizens. Data-processing ADM systems provide opportunities to increase the amounts of information on which decisions are based. These information sources contain personal information. Examples are data about loans, education and tax payments. The act of translating these information sources into inputs for decisions or decisions themselves is often externalised to intermediaries within the organisation, to increase efficiency by task division of policy, implementation and supervision (van der Linde, 2021). Using AI in the public spheres then leads to increasing amounts of citizen information used and intermediary parties which process that citizen information. If citizens do not have oversight over this extra information and data processing intermediaries, there is a vulnerability of increasing information asymmetry about social decisions between citizens and governments. Richardson et al. (2019) even call this a power imbalance.

4.5.2 Panel VIII: Societal, After implementation

Reward Hacking
Some reward functions of algorithmic systems or self-learning systems like AI are prone to gaming or tricking of their reward function, i.e. the function which tells the system what to act upon (Hadfield-Menell, Milli, Abbeel, Russell, & Dragan, 2017). Citizens or users within governments can show behaviour which suits the reward function of the AI model, but is not intended by the developers or implementer of the model. This is often strategic behaviour (Dobbe & Raji, 2019). This is a vulnerability as it potentially leads to harmful situations for citizens whilst not getting stopped because the AISs seems to function properly. The AIS is intended to serve some public goal, but brings unintended harm due to the reward hacking by others.

Unexpected or unwanted societal feedback loops
AISs in societal contexts can bring about changes to this societal context itself. If the AI model subsequently gathers input data from this altered societal context, a reinforcing feedback loop between the societal context and the AI model potentially occurs. Dobbe & Raji (2019) mention policy shifts as on of these unwanted feedback loops. The input variables for the model are altered by the policy which is the output of the model, and the feedback mechanism can offset the intended positive societal effects of the model or even cause harm to the citizens it was ought to serve. A classic example of such a reinforcement loop is found in predictive policing models. Police officers are sent to over-policed areas in cities based on historical arrest data, again arrest people in these areas and subsequently feed new arrest data which overly focuses on these same areas into the model. Richardson, Schultz & Southerland (2019, p.23) in this context refer to "collateral consequences" of ADM systems. If ADM systems in for example criminal cases make incorrect decisions about citizens, this unjust criminal conviction is a potential reason not to receive governmental benefits. Citizens might go downhill in other aspects than the model was intended for, hence the designation 'collateral'. Unpredictable human interaction with the model further complicates the ethical question of how and when to deploy the system for the first time (Dobbe & Raji, 2019).

Developments in legal control and liability over decisions about citizens
Governments use AISs to inform their own choices or make choices for them. If decisions are made by the AIS, these decisions impact the lives of citizens to some extent. Citizens have the right to check such public decisions for lawfulness and governments must be held liable for their decisions about citizens' individual lives. The question then rises how governments can be held liable for decisions made by AISs to make such decisions (Wirtz, Weyerer, & Sturm, 2020), and citizens can exercise their democratic and legal control. This creates a vulnerability for citizens in terms

of democratic control over public AI-fueled decision-making. If the working of AISs is unknown to the subject of decisions due to intellectual property rights, this matter of liability becomes even more tense. The AI Now Institute (AI Now Institute, 2018) therefore opposes the use of AI models which have trade secrets in the context of decisions about benefits for citizens. This discussion gets even more exciting due to the increasing autonomy and therefore agency of AISs (Gervais, 2020).

Contribution to existing social disparities
AI developers use data to train, test, validate and use models. Section 4.2 describes several vulnerabilities related to the biases that data sets needed for these modeling steps potentially contain. Additional to this prior knowledge, it is striking that ADM systems are often used for fraud and abuse detection. Richardson, Schultz & Southerland (2019, p.21) even speak of a "national trend to target poor people under the auspices of prosecuting “waste, fraud, and abuse” in government systems". This focus on fraud leads to a focus on citizens in need of social benefits. That is because governmental data sets already contain their information, leading people with less money to be in the spotlights of AISs like these. Beneficiaries are more precisely monitored and checked than taxpayers (van der Linde, 2021). Considering the biases in data and focus on fraud and abuse detection, there is a vulnerability that AISs contribute to existing inequalities between socioeconomic subgroups in society.

4.5.3 Illustration of harms due to societal vulnerabilities

A case from New Orleans strikingly illustrates the vulnerabilities concerning the lack of societal control on governmental AI usage. The city of New Orleans used an automated network-generating system based on social media connections to assess whether people were likely to become either a victim or offender of gun violence. The purpose of the system was to help law enforcement to intervene and offer support to people likely to show criminal behaviour (Richardson, Schultz, & Southerland, 2019).

Later, a man was sentenced to 100 years in prison based on alleged criminal conspiracy. His alleged ties to other gang members were a focal point in this court trial. His lawyers learned about the system only after the conviction and wanted to re-open the case and use the network materials. This appeal was denied. Prosecutors said the system did not play a role in the case (Richardson, Schultz, & Southerland, 2019). This case points to the vulnerabilities of increasing information asymmetry due to opaque technological developments and questions about legal control over governments. Besides, it also questionable what is a rightful threshold of a person’s social network to perceive him or her to be prone to criminal perpetration or victimisation.

In February 2020, a Dutch court declared fraud detection system SyRI to be unlawful. This system combined data concerning many different socioeconomic factors about individuals, for example about past detentions and education level, to create risk profiles for fraud conduction. SyRI only created risk profiles in less advantaged areas for some larger municipalities in the Netherlands. The court decided that using SyRI was illegitimate (Huisman, 2020). One of the reasons was lack of transparency for citizens, which makes citizens vulnerable to lacking democratic control. The court also mentioned unwanted stigmatising and discriminatory aspects of the system, as it was only used in so-called underdeveloped city areas. This relates to the vulnerability of contribution to existing societal disparities in this model layer of societal context of vulnerabilities.

4.6 Chapter Conclusion

This chapter for stage 2 of the research aimed to answer the following sub-question:

Seen from an AI system perspective, from which relevant contexts do vulnerabilities originate and how can these be structured in a conceptual model?

This thesis takes a system approach, see section 1.2.2. It does so to not only consider the AI technologies but interpret it as part of a group of connected technological and non-technological subsystems and the multi-actor network involved with it, stemming from the conviction that only this leads to nuanced and relevant answering of the main research question. The system perspective reflects throughout this entire chapter. Technological vulnerabilities are vulnerabilities from multiple technological subsystems, for example the required data for model development, but also the security of the technological infrastructure, as data sets with sensitive information lead to a vulnerability of leakages. Non-technological subsystems are well represented as well, for example when legal vulnerabilities are considered. The multi-actor component is touched upon as the involved actors are all represented in the model contexts of the AIS: from model developers to the citizens of Amsterdam.

The 'onion' representing the four contexts has an inherent logic (set out in section 4.1.1) which makes it a useful conceptual model to use for vulnerability interpretation. Four relevant contexts to consider are: *AI model*, *Model deployment*, *Political administrative*, and *Societal*. Vulnerabilities as presented in this chapter from all of these contexts do not give an exhaustive overview, but the contexts taken together cover the scope of the issues governments possibly encounter. Policy makers can use this model to understand the different context-dependent vulnerabilities of their AISs. It helps them to set out directives to think about the potential risks of their new system, to from there on come up with governance requirements and mitigation measures for different involved teams from the four contexts. The next step for this research is to check whether this generic model is valid to use by comparing real-life AIS vulnerabilities found in the CoA cases with the model vulnerabilities. This comparison also sets directions for complementing of the model.

Chapter 5

Extended Case Descriptions and Model Validation

After having completed the literature review, there are ex ante ideas about the research variables. The next step is to enter the field and to collect data (Eisenhardt, 1989). The research thus enters stage 3 "Understanding the cases, validate and complement the vulnerabilities model". This chapter presents relevant results from the interviews to extensively describe the interesting case aspects and to validate and complement chapter 4's conceptual model. First, basic information about the interviews is presented.

5.1 Basic Information Interviews

In appendix B, table B.1 presents generic information about the nine interviews for three cases. Appendix C contains the questions for the interviews. The interviewee names are not disclosed due to privacy reasons. The selection of interviewees is based on the literature review and availability of municipal employees. In chapter 4, a conceptual model to understand vulnerabilities of public sector AISs with different contexts is presented. The different layers in this conceptual understanding are: the AI model itself, the model deployment, the political-administrative layer and society. To thoroughly understand vulnerabilities, stakeholders from these different contexts have to be interviewed. For every case, public agents with expertise from the layers *AI model*, *model deployment* and *political-administrative* are selected. Due to time and practical reasons in this Covid-19 pandemic, no interviews with citizens from the *Societal* contexts were held. Their perspective is taken into consideration by explicitly asking every interviewee how the citizen perspective played a role in their work for the AIS.

Based on the interviewee's job and expertise description, their specific role or expertise is considered to be one of these three contexts. However, this does not mean that the generated data from every interview can only be interpreted within the contexts itself. For example, an AI model developer can provide useful information about societal vulnerabilities inflamed by AISs as well.

The subdivision of interviewees over the above mentioned conceptual model contexts is made as follows. Public agents who carry or carried ultimate responsibility to the political and administrative actors for the full or partial functioning of the department for which the AIS was used, are assigned to the *Political-administrative* context. Interviewees who are not responsible for the functioning of the system, but are involved and have in

interest in the well-functioning of the AIS to fulfill their task in the specific application domains, belong to the *Model deployment* context. Lastly the people who actually technologically contributed to the functioning of the algorithmic systems considered in the three cases, are assigned to the *AI model* context.

5.2 Expanded Case Descriptions and Interviewee Involvement

Section 2.5.1 presented short descriptions of the three cases which are considered for this research. This section contains more wide-ranging descriptions of the cases, based on the interviewee's answers. First, this section presents a tabular cross-case overview with relevant case information in table 5.1. The following subsections provide the longer textual descriptions. Furthermore, every expanded case descriptions concludes with the job and expertise descriptions of the interviewed workers who were or are involved with every AIS. These descriptions contain no exact function titles or personal pronouns for privacy reasons.

Table 5.1: Tabular overview of relevant case descriptors

	Case: Reporting issues in public space	**Case: Automated parking control**	**Case: Illegal holiday rental housing risk**
Involved departments	Research, Information & Statistics (OIS), Amsterdam Service Center	Parking Services	Housing, Surveillance & Enforcement
Case-specific goal of departments	Resolve public space issues as quickly as possible	Increase enforcement effectiveness and efficiency	Increase enforcement effectiveness
AI model contribution	Categorise citizen reports automatically	Recognise license plates on a large scale at fast pace	Prioritise complaints for human agents to check for, based on risk estimation
AI technology	Machine Learning	Machine Learning	Machine Learning
Model architecture	Logistic Regression	Unknown	Random Forest Regression
Model type	Text classifier - language processing	Image Recognition	Ensemble learning
Input data	Reports filed by citizens	Scanned license plates	Complaints, house data, house owner data
Outsourcing partners involved	-	Egis Parking Services	-

5.2.1 Case description: Reporting issues in public space

People who encounter problems like noise pollution, cafe nuisance, odour nuisance or litter in Amsterdam, have the possibility to report these problems. The municipality used an old IT system, Kimora, which received the reports and had to forward the report to the right authority within Amsterdam, which has the responsibility to solve the issue. Citizens had to choose from a very large amount of nuisance categories themselves, which made the report form very unclear and the process of filling in a report online unnecessarily lengthy. Or as put in I1: *"Eskimo have one hundred words for snow, we have one hundred categories to report issues in public space."*

The enormous number of categories to choose from let many citizens to interrupt and not fulfill their report. Furthermore, many people consciously or unconsciously chose

the wrong nuisance category. This means that many of the reports had to manually be placed in the right category by municipality workers, leading to higher handling times and some issues not being resolved. As much as 80% of the service requests was not handled in due time.

Another important problem was that the IT system, Kimora, was old and outdated. Due to the high number of reports it had to handle, it was bursting it banks. A renewed system that was easier to make requests for citizens and more stable was necessary. At the time the number of blackouts and delays of the system was getting really problematic, an AI developer (I2) presented a relatively simple supervised ML model for automatic text classification. Municipal agents were pleased with the opportunities it brought and this simple ML model was used as the basis for a completely new system for handling citizen reports: SIA.

SIA last year handled over 380.000 service requests and the number of requests increases by approximately 20% every year. Citizens fill in the issue they encounter in a free text field online, and the ML model predicts the right category and stalls the report with the right municipality department to resolve the issue. Human agents check whether reports are correctly categorised, as well as empty the 'others' category where issues are stalled for which the ML model did not have enough certainty (40%) to classify the text and categorise the report.

The interviewees are:

- **I1: Political-administrative context.** Responsible for the program which aimed at improving the processing of citizens' public space issue reports. Lead the "Actie Service Centrum", a newly established municipal department responsible for managing the incoming reports. Decided to implement the ML text classifying model to categorise reports.
- **I2: AI model context.** Developed the ML model used for automated text classification. Was kept involved with the program for some time after the model was implemented, but not anymore. Development operations are now assigned to others.
- **I3: Model deployment context.** Activities concerning the public space issue reports are twofold. Firstly, involved with the actual resolving of the public space issues by selecting the right people to resolve and coordinating if partnerships between the different resolving departments within the CoA. Secondly, later became responsible for the complete range of nuisance categories which can be assigned to citizen reports.

5.2.2 Case description: Automated parking control

The City of Amsterdam wants to be a low-car traffic city and stimulate the use of public transport, bicycling and walking as transport means. This goal is one of the reasons to set high parking tariffs and aim for effective control of parking payments. Citizens who are checked and have no permission to park their car on a certain parking spot, either trough a long-term or short-term parking license, receive a fine. To increase control capacity and save money, the municipality have been using scanning cars to check parked cars for parking licenses for some time now.

The scanning car system is subcontracted to an external parking company, Egis Parking Services. Egis deploys the cars and checks for legality of parking. The system works as follows. Scanning cars scan license plates and take photos of cars' surroundings. A ML model uses image recognition to determine the license plates of the cars scanned. Based on this license plate scan, the parking license is checked for in a countrywide central database from the Dutch vehicle authority *RDW*.

If a car is eligible for a fine due to a lacking parking permission, a human agent checks the surroundings photos taken by the scanning to see whether there are circumstances which clear the illegally parked car, for example if the car only stopped for loading and unloading. Amsterdam's parking department does not check the ADM systems themselves, and only steers for contractual KPIs like level of payments of scanned cars. Another relevant aspect is that the ML model for license plate recognition is provided to Egis Parking Services by another external developer. This developer is subcontracted by Egis Parking Services, not the CoA.

The interviewees are:

- **I4: Political-administrative context.** Manages the parking department and is responsible for street parking enforcement. In that capacity, also ultimately responsible for the contract and contact with external contractor Egis Parking Services. Must carry out the parking policy and street parking enforcement policies established by the bench of Mayor and Aldermen and city council.
- **I5: Model deployment context & AI model context.** This interview was conducted with two interviewees. The first interviewee works for the municipality and manages the contract with Egis Parking Services. The interviewee therefore is in close contact with the Egis technological experts and also determines the KPIs based on which Egis' performance is evaluated. The second interviewee works for Egis and is responsible for analysis of data generated with the scanning cars and development operations of the algorithmic systems developed by Egis, which e.g. compare the scanned license plates with parking permission registers.
- **I6: Model deployment context.** Uses data generated by scanning cars to report about street parking in Amsterdam. These insights are used for future decision-making and street parking policy.

5.2.3 Case Description: Illegal holiday rental housing risk

Many house owners in Amsterdam rent out their homes, or parts of their homes, to tourists visiting the city. Airbnb is the most used platform for these rentals. To arrange overnight stays for tourists in such way brings benefits to both house owners and tourists, but also has downsides for the city. Think of tourist overloads and rising house prices due to investors who buy houses solely for holiday rental. To limit these holiday rentals, the municipality decided to set a yearly maximum of 30 nights of private rental per house and a maximum of 4 tourists per rental.

Citizens in Amsterdam who think that others do not obey these rules, can open complaints with the municipality. Municipal workers then decide whether or not to check the house about which complaints are opened. House owners who indeed illegally rent our their house potentially receive fines up to 20,000 euros. To improve enforcement performance by increasing the percentage of illegal rentals per houses checked, the CoA have decided to pilot an AI model which prioritises the illegal rental complaints. Due to Covid-19 travel restrictions in 2020 and 2021, the number of tourists in Amsterdam is too low for the pilot to be relevant and therefore has not yet started.

The model is a Random Forest regression model which classifies the complaints based on information about the house, house owner and complaint itself. Data sets that are used e.g. concern the square meters of the house, number of Airbnb reviews for the house, personal information about the house owner and the time at which the complaint was opened.

Based on the model classification of complaints, a hierarchy for human agents to check complaints in a certain order is created. This hierarchy again aims at increasing the

chance of catching illegal rents. For every case, the human municipal workers who check the complaint are informed about the data which made the AI model assign a certain urgency for checking the house. The human enforcement workers do not have to follow the exact order created by the model, as they are allowed to check other complaints based on e.g. own insights or practical routing reasons.

The interviewees are:

- **I7: Political-administrative context**. Manages the team ("werkplaats" in Dutch) which develops the AI model and carries responsibility to the Management Team and Alderman for this policy for illegal holiday rental enforcement.
- **I8: AI model context**. Works as a data scientist and part of the technological development team which created and improves the AI model for prioritising the illegal rental complaints. For example works on measuring feature importance of input data for the model.
- **I9: Model deployment context**. Legal expert who is involved with the team responsible for the enforcement of illegal housing rentals. As a legal expert, the interviewee contributes to legal compliance of the model. Furthermore the interviewee works on the substantive reflection of how the enforcement team should operate, for example through assessing business rules of the team.

5.3 Validation Context-Dependent Vulnerabilities Model

The validation of the vulnerabilities model is based on two parts. First, an argument is made based on the used literature for the model. Table D.1 (appendix D) presents all articles and reports used to argue for the conceptual model. The model relies on a large amount of (semi-)literature sources, 41 in total. Furthermore, the collection of sources shows a great variety of themes which lays the groundwork for a nuanced and valid-to-use model:

- 5x Computer Science
- 3x Computer Science & Ethics
- 6x Computer Science & Law
- 6x Computer Science & Law, Computer Science & Public Administration
- 1x Computer Science & Philosophy
- 6x Computer Science & Public Administration
- 9x Computer Science & Society
- 1x Computer Science & Society, Computer Science & Ethics
- 3x Data Science
- 1x System Science

The large and strongly varying amount of sources is a first positive note on the validity of the vulnerabilities model.

One of the goals the conducted interviews serve is to partly validate the conceptual model of vulnerabilities as presented in chapter 4. Interviewees were asked for aspects of the AIS which they see as downsides for citizens. Some of the downsides were not explicitly labeled as negative aspects, but can be interpreted as such. Using the term vulnerabilities during the interviews was avoided because this might lead to confusion or having conceptual discussions rather than focusing on substantive problems for citizens caused by the three AISs.

Qualitatively validating the conceptual model of AIS vulnerabilities then works as follows. All interviewees came up with aspects of the AISs which they considered to be downsides. Only the downsides that directly or indirectly affect citizens and not only the municipality as an organisation are relevant for this research. First, table D.2,

table D.3, and table D.4 in appendix D present all these mentioned negative aspects. An appendix is used due to the size of the tables and large amount of negative aspects per AIS. The tables also tell in which interviews the aspects were discussed. Lastly, the final column of the tables assign so-called Aspect IDs to the downsides for the comprehensibility of the next part of this validation.

The validity of the conceptual model is acceptable if a large part of the case-specific downsides mentioned by interviewees fit the vulnerabilities of the model. If some of the negative aspects are impossible to interpret as one of these vulnerabilities, these point towards reasonable adjustment and complements for the conceptual model. All contexts - *AI model*, *Model deployment*, *Political-administrative*, *Societal* - contained vulnerabilities. For every mentioned downside by interviewees, now referred to with the Aspect IDs, are interpreted as belonging to one of the four contexts if possible. Next, they are assigned to a specific vulnerability if possible. The assignment of the case-specific downsides mentioned by interviewees to generic public AIS vulnerabilities is presented in table D.5, table D.6, and table D.4 in appendix D.

Table 5.2 presents an overview of the conducted validation. It contains the total number of mentioned downsides per studied case, the number of downsides that were interpretable as vulnerability using the conceptual model, and the specific aspects which did not fit the vulnerabilities model. For every studied case, an acceptable share of the downsides of the AISs mentioned by interviewees fits into vulnerabilities model. The mentioned aspects which do not fit the model, point towards adjustments or complements for the model. The closing section 5.5 of this case study results chapter discusses these adjustments and complements.

Table 5.2: Overview model validation

	Case-specific vulnerabilities	**Interpretable using model**	**Not interpretable using vulnerabilities model**
Reporting issues in public space (RI)	20	17	RI.2 Certain need for AI system which was never intended to use for large-scale municipal operations like these, concerns about controllability RI.7 Human agents blame the AI model for mistakes they wittingly or unwittingly make RI.17 Sudden interferences with priorities for system development from alderman, which serves her goals but might not be of interest for the complete population
Automated parking control (PC)	14	12	PC.4 Data leakages have occurred PC.7 Lacking capacity to do surroundings photo checks for all scanned cars without parking permission
Illegal holiday rental housing risk (HR)	18	17	HR.16 Team members have access to data sources to which they should not have access

5.4 Interpretation of Notable Case Study Results

This section presents interpretations of the vulnerabilities found - and vulnerabilities not found - in the case studies.

Reporting issues in public space
Three interviewees for this case in total only mentioned one vulnerability which can be interpreted as a political-administrative one: "Lack of development experts which leads to compromises on AI system quality and strong dependency on individual experts". Seen from the high number of societal vulnerabilities mentioned, it is remarkable that no interviewee mentioned a downside which can be interpreted as the political-administrative vulnerability "Relying on AI models undermines broader political debates". This AIS provides citizens with the opportunity to influence their own public space, and the system functions satisfactory, but at the same time does seem to create a new bias regarding certain subgroups of citizens who are e.g. more outspoken or have higher trust in government. Using AI for efficiency of operations in this way possibly undermines a broader political debate about how the CoA should manage its public space both objectively and effectively. This is especially remarkable because the interviewees did mention the societal vulnerabilities themselves: the next step is to understand that the extent to which the CoA lets the AIS influence the public space is a political choice as well.

Automated parking control
Two notable vulnerabilities for this case are "Dependency on outsourcing partners for technology development - no oversight from municipal agents" and "Outsourcing partner determines car routing and surroundings checks - no steering from municipality apart from KPIs and checklists". These vulnerabilities are both interpreted as chapter 4's model vulnerability "Shifts in discretionary power public agents" in the political-administrative context. Shifting discretionary power thus becomes a striking vulnerability when an important share of the AIS is outsourced to an external party, in this case Egis Parking Services. This phenomenon is also mentioned by others, e.g. (Richardson, Schultz, & Southerland, 2019). The involved political-administrative actors should therefore be extra careful to discuss which discretionary power is assigned to external developers and deploying actors in AISs like these. In this case, some interviewees raised questions whether the external deploying company had the right financial incentives for their parking car routing, or it should be stimulated to focus on fair routing in terms of touching upon every area in the city rather than receiving financial benefits through optimising willingness-to-pay.

Illegal holiday rental housing risk
The ML model for this AIS determines the chance that a complaint about Airbnb rental fraud is rightfully made based on many different data inputs from multiple sources. These kind of fraud detection models are under scrutiny because of the *Toeslagenaffaire* and the SyRI case. It is therefore remarkable that many vulnerabilities mentioned by the interviewees are "Algorithmic system makes classifications based on business rules which are not formally determined or described", "Draw conclusions about housing fraud based on input information that might not completely be relevant for this outcome", "Hard to detect bias because it is hard to define the relevant societal subgroups in real life" and "No substantive arguments for all relationships between input factors and model outcomes". These are all related vulnerabilities which state that the fraud detection model sometimes lacks the right reasons to do what it does. It points to a discussion which is particularly interesting with AISs like these which use large amounts of data. The question is whether relationships determined by BD models are to be used by governments if these are hard to check or argue for based on for example real-life domain knowledge or legal foundations.

5.5 Chapter Conclusion

This chapter is centered around research phase 3 and sub-question 4:

What do the case-specific vulnerabilities found in the City of Amsterdam's practice say about the validity and directions for complementing of the conceptual vulnerabilities model?

The first method to assess the validity of the vulnerabilities model is based on an overview of the literature used to construct it. 41 different sources, varying from scientific articles to expert reports to newspaper articles, together contributed to the construction of the model. These sources have eight different themes, for example system science or computer science & ethics, or some combinations of themes. Using such a large quantity of sources from different kinds and concerning different themes is not enough to prove the validity of the vulnerabilities model, but it does prove that the model was not constructed overnight.

More important is the assessment of the vulnerabilities model based on the case study lessons from real-life. After all, the knowledge gap presented in section 1.3 states that there is a need for relevant theorising of public sector AI use, theory which consequently aids in addressing the knowledge gap of AIS risk management for governments. The theoretical construct about vulnerabilities is valid and therefore relevant to use later this thesis if the case-specific downsides of AISs in Amsterdam are interpretable using the model.

The expert interviews for the three case studies in total yielded 52 case-specific vulnerabilities of the AISs. 6 of them did not fall within one of the vulnerabilities in the model contexts. This is an acceptably low share, and combined with the literature-based lead as just described, the model is considered to be valid and further usable. The next step hence is to use the vulnerabilities model to come up with requirements and vulnerability mitigation strategies for public sector governance of AIS.

The last concern of this chapter is to find complements for the model based on vulnerabilities found in the CoA's practice that did not fit into the literature-based generic model. The 'unplacables'[1] from table 5.2 point towards important complements for the vulnerabilities model:

- A vulnerability of the *model being adopted without enough capacity to control it*. This addition is based on RI.2 and PC.7. The vulnerability here is that the model brings a need for control when implemented, for example devops or human oversight over decisions. If this capacity for control is underestimated before the model is implemented, this brings the danger of malfunctioning systems or unchecked model decisions which should have been checked. This addition is related to the political-administrative vulnerability "Inadequate financial resources allocated to AI development teams", as this can be the cause for the lacking capacity for control. Nevertheless, this complement still has added value as it points towards the new and extra forms of capacity necessary for AISs which are often overlooked upfront.
- A vulnerability concerning *inadequacies in the requisite security of the AI model*, based on PC.4 and HR.16. Governments should be aware that AI models often require sensitive data sets and that these should be protected. So care should not only be taken of quality measures for the model itself, but also of the surrounding infrastructure, so that no personal citizen information falls into the hands of municipal employees or criminals who should not be able to have it.

[1] This thesis' English language enrichment and a nod to one of De Niro's great films: The Untouchables

- A vulnerability of *unwanted strategic behaviour with the AI model*. This directive for model improvement is based on RI.7 and RI.17. The adoption of AISs provides model deployment actors (RI.7) and political actors (RI.17) with new possibilities for strategic behaviour. If this behaviour comes at cost of the quality of the decision-making, which was the case here, the models bring new vulnerabilities which might result in harms for citizens. This complement relates to the societal vulnerability "Reward hacking", which can also be strategic behaviour. However, reward hacking is to purposely game the outcomes of a system to use it for your own benefits at the cost of others'. This complement refers not the gaming of system outcomes, but using the newly adopted AIS in an unwanted way other than manipulating its outcomes.

Chapter 6

AI Governance Requirements

This chapter reports about the prescriptive stage 4 of this thesis: "Finding governance requirements and strategies to deal with vulnerabilities". Chapter 5 validates the usefulness of context-dependent consideration of vulnerabilities of public sector AISs. Because the validity of the vulnerabilities is acceptable, the model is also used for prescriptive purposes in this chapter.

A second scientific knowledge gap from section 1.3 is the lack of knowledge of effective governance to deal with risks of using AI in the public sector. This chapter relates to this incomplete knowledge. It presents requirements and strategies for meticulous governance of public sector AISs, based on relevant requirements and strategies discussed during the interviews with experts contributing to AISs in the CoA. Several vulnerabilities from the conceptual model are linked to governance requirements and mitigation measures found in practice.

These requirements are presented in sets: dilemmas. The dilemmas contain two governance requirements that generate a certain degree of friction: if the CoA commits to one of the requirements in the set, the other could be compromised. This idea became apparent during the case studies: it is not interesting to only translate vulnerabilities into requirements. Otherwise it seems as if the challenges could all simply be mitigated step-by-step. The idea that governance choices in the field of public sector AI sometimes come at cost of each other, is also found in literature and therefore introduced in section 1.2.5.

To start the chapter, the next section defines and distinguishes governance requirements and mitigation measures.

6.1 Defining Requirements and Mitigation Measures

In the context of AISs, requirements can have several meanings. Veale & Brass (2019) examine requirements for bureaucratic control over public sector algorithmic systems through a juridical lens: to what extent do these systems meet existing soft or binding laws. Others refer to ethical requirements that public sector organisations need to conform to when using AI in their operations (Wirtz et al., 2019).

This chapter is aimed at finding relevant requirements for public sector organisations using AI. Again, governance here is in essence defined as: "Governance as steering and coordinating"[1]. An important aspect of governance for public sector AISs, is the fact

[1]For an elaboration on this definition, see section 1.2.3

that trade-offs are inherent to it, as set out in section 1.2.5. It is more interesting to find trade-offs found in Amsterdam's practice than to only sum up singular requirements, because these would be quite intuitive and would not do justice to the complex public sector reality. Lastly, section 1.2.2 introduces the system perspective as analytical lens for this thesis. The interdependencies between the four different contexts as well as the involved actors are just as interesting to analyse as the mere technological aspects of public sector AI. Pooling the above mentioned definitions, notions and analytical lenses brings us to the following interpretation of public sector AI systems requirements:

Definition public sector AI systems governance requirements:
The needs for coordination and steering of which governments must be aware when they seek to address interdependent vulnerabilities of their AI systems.

Mitigation measures are measures aimed at reducing the harmfulness or unpleasantness of something. For the public sector, the Dutch governmental auditor Algemene Rekenkamer (2021) calls for an interdisciplinary way of mitigating the risks of public sector algorithmic systems like AI, using multiple sorts of mitigation. This thesis follows that recommendation. The multiple types of measures relevant to mitigating the downsides of public sector AI are presented in section 3.3:

- Technological measures
- Data measures
- Monitoring and Evaluation
- Legal measures
- Organisational measures
- User engagement and citizen agency

Combining these notions, the following definition of mitigation measures for this research follows:

Definition mitigation measure for public sector AI system:
Technological or governance measures to prevent or reduce the harmful consequences of public sector AI systems' vulnerabilities.

In the next sessions, all governance requirements are along-sided by the relevant mitigation measures from different categories found in the case studies.

6.2 Requirements Dilemma I: Increasing and Decreasing AI's Impact

This section presents the first set of governance requirements based on the vulnerabilities model and found during the case study.

Observation 1: increased impact to justify using intrusive models

In the Illegal housing rental case, the AIS serves to better prioritise complaints about potential illegal Airbnb housing rentals by fellow citizens in Amsterdam. This prioritisation is conducted using a model which should outperform humans in terms of pattern recognition and calculate chances of valid complaints, and hence increase efficiency and effectiveness (see section 3.1 for an elaboration on these merits of public sector AI) of the CoA's enforcement on Airbnb fraud. The model needs personal

information about citizens, for example information on households and the house itself. To rationalise the intrusive character of using such a model, the eventual impact of the model on the decision-making process should be significant and thus allow the efficiency and effectiveness increase. Such significant impact could be to replace human enforcement capacity with model decisions completely. The following quote reflects the idea of one of the interviewees who wants to increase the impact of the algorithmic system on decision-making:

"Now we have to switch to some proactive, predictive algorithm, for which addresses where no complaint is made, can also play a role" - Interviewee I7

Observation 2: decreased impact to justify using a model which has its flaws

Again, in the Illegal Housing Rental case, the interviewees sought to minimise the impact of model decisions as well, because due to the abovementioned intrusive character of ADM based on personal information, the CoA employees did not want to completely make decisions based on the model and ensure human oversight. Furthermore, the potential mistakes of model decisions can have severe harms because citizens might receive fines up to thousands of euros. The following quote reflects the opinion of one of the interviewees, who thought that only using the model to prioritise complaints rather than making the decisions itself was a good idea to overcome challenges:

"What we are doing is (...) prioritizing the cases, all the cases are going to be tackled the only thing that changes is the order that you will tackle the cases. The model in this case serves the purpose for what is needed so in this case is fine.." - Interviewee I8

Governance requirements I: dilemma

The observations point to the following governance dilemma. First, if a public sector organisation decides to use algorithmic systems, especially ones that could be politically or socially sensitive, or are intrusive and use personal citizen data, the system should at least have a thorough impact:

Requirement I-a:
Increase the impact of the AI model on the decision-making process to ensure its added value and create balance with the downsides of the AI system.

At the same time, potential model mistakes must be compensated and the impacts of erroneous model decisions must be alleviated. First, governments like the CoA must determine what is an acceptable effect of model errors, and the impact of the model on decision-making must be neutralised in order for this acceptable margin of error not to be exceeded:

Requirement I-b:
Decrease the impact of the AI model on the decision-making process until the effects of model errors are acceptable or errors are still possible to mitigate.

AI System vulnerabilities leading to requirements dilemma I

This governance dilemma results from the CoA seeking to address different advantages and vulnerabilities of public sector AISs. For governance requirement I - a:
Vulnerabilities - Political-administrative context: "Weak relationship between AI deployment and public policy goals".

Advantages: Efficiency - "Addition to and better utilisation of existing human capacity". Effectiveness - "Process larger amounts of information in times where much data is available", "Learning system outperform humans with pattern recognition", "Better aimed use of capacity". Benefits for citizens - "Potentially more objective decisions without partiality", "Consistency of decisions".

The vulnerabilities which lead to governance requirement I - b are:
Vulnerabilities - AI Model context: "Mathematical accuracy of decisions", "Inadequate fail safe mechanisms and plan B procedures", "Different accuracy of decisions for different subgroups population".

Mitigation measures for requirements I-a and I-b

Several useful mitigation measures, categorised based on the categories from section 3.3, were mentioned to cautiously contribute to these two governance requirements. For governance requirement I-a, these are:

- **Organisational**: Assign an extra organisational department (like the "Actie Service Centrum' from Reporting Issues) with the task to control the AIS and oversee the decisions' quality
- **Organisational**: Ensure mandate for increasing impact of the AI model by involving the city council.
- **Data**: Provide development teams with enough labeled data to compare the old system performance with the new system performance whilst the model is in use.

For governance requirement I-b, several mitigation measures to prevent mistakes or compensate for mistakes made by models were found in the case study:

- **Technological**: Use a threshold of certainty for the model to make decisions.
- **Data**: Check whether the training data sets contain biases which influence the accuracy of model decisions for societal subgroups - coined bias analysis in the Illegal housing rental case.
- **Monitoring and evaluation**: Have AIS decisions checked by a human employee and corrected if necessary.
- **Monitoring and evaluation**: Check for individual circumstances that clear citizens of guilt, as seen with the human image review in the Automated Parking Control.
- **Legal**: Make sure individual model decisions are traceable so legal actors who want to check model decisions for legality have the opportunity to do so.
- **User engagement and citizen agency**: make sure the model only advices the human enforcement agents by e.g. prioritising rather than taking decisions in the Illegal Housing Rental case.

6.3 Requirements Dilemma II: Freedom and Restrictions for Developers

This second governance requirements section also contains a dilemma of two requirements, which together result in a tension.

Observation 1: modelers have limited time and must be given freedom to create high-end technology

This observation holds for all three cases. If developers, often short in time anyway due to their unique skills, have to spend too much time on peripheral matters and are tied down with rules, governmental departments will not be able to reap the full

benefits of AISs. One interviewee luckily had enough time to work on technological solutions rather than on other issues that appeared within the 'business':

"I try to avoid every kind of business, I'm really a tech person. I like to talk with people and everything but not to deal with business problems that's definitely too much" - Interviewee I8

AI development simply requires skills that only few people have and they should be allowed to take decisions about the model development based on their expertise:

"The question is: do we still have the know-how to make such judgements? I don't know. At least it is not in the places where the decisions are made." - Interviewee I2

If a government decides that it wants to reap the benefits of AISs as presented earlier this thesis, e.g. because it outperforms its human employees for pattern recognition, it cannot stop the development in its tracks.

Observation 2: models have to be fit into the daily governmental operations and employees must accept its use

In the Reporting issues case, the new text classifying ML model in the beginning had fewer functionalities than the old system, which led to unrest within the organisation:

"Sometimes we encountered user groups who stood in front of the model developer with pitchforks and torches, saying "we want new buttons!", then I had to tell them that the new buttons were coming." - Interviewee I1

In the Illegal Housing Rental case, an interviewee points towards the importance of explainability of the model as well:

"You need to make everything super explainable, even to non-technical people" - Interviewee I8

It comes down to the following: AI models can only be of value for the organisation if the lion's share of the employees who actually fulfill the decision-making task both understand and accept the functionalities of the model. Modelers can therefore not just focus on the technological quality of their products, but must also think of the functionalities for people working with it and explainability of the model's functioning as well.

Governance requirements II: dilemma

The first requirement concerns that governmental decision-makers must allow freedom for developers to have enough time to work on their technology development and let them create high-end technological products to their own professional discretion:

Requirement II-a:
Leave model developers with enough time and professional freedom to create high-end technological products.

A second governance requirement which is at odds with requirement II-a, represents the observation that the AIS must be aligned with the existing organisational functioning and employees in order to be adopted and actually used:

Requirement II-b:
Ensure that developers create models which are explainable and functional for governmental employees to be used in their daily practice.

AI System vulnerabilities leading to requirements dilemma II

The AI model contextual layer of chapter 4's vulnerabilities model contained many technological difficulties which emerge before the model is implemented. These vulnerabilities must be dealt with by highly-skilled technological experts, who must have freedom of action to do so. Therefore these following vulnerabilities lead to the governance requirement II-a:
Vulnerabilities - AI Model context: "Overfitting and underfitting", "Poor data quality", "Training data gaps", and "Technological zero-sum games and trade-offs during modeling".

Within the model deployment context, vulnerabilities of AISs emerge which concern the workers and addressing these vulnerabilities leads to governance requirement II-b:
Vulnerabilities - Model deployment context: "Lack of model understanding with deploying department's workers" and "negative impact on workforce".

Mitigation measures for requirements II-a and II-b

Mitigation measures found in the case studies which contribute to the fulfillment of requirement II-a are:

- **Technological**: Use algorithmic methods, like the so-called SHAP method[2] in the Illegal housing rental case, which determines which input information led to the model decision. In this way you can use automated methods to improve the understanding of model deploying employees.
- **Organisational**: Have multi-disciplinary development teams in order to divide the tasks based on expertise of team members. In this way developers are not responsible for the communication about the model, reporting the steps taken and so forth.
- **Monitoring and evaluation**: Make sure there are ongoing evaluation sessions between developers, principals and model users. In this way not everything has to be reported and laid down, which takes more time. It also increases the quality of cooperation between developers and domain users, which ensures a better alignment of the model with the deployment domain reality.

Mitigation measures for governance requirement II-b:

- **Technological**: Choose a model architecture, in the Illegal housing rental case a Random Tree Forest, which might be outperformed in terms of accuracy by some more advanced modeling techniques, but is more explainable to non-technological experts and citizens.
- **User engagement and citizen agency**: The inputs that led to a model outcome must be shared with human workers, so that they understand what information led to the model decision and they can shine their own light on that decision.
- **User engagement and citizen agency**: Create a list of the explicit and implicit business rules which determine the functioning of the model. Let these business rules be formally accepted by the model deploying domain experts.
- **Monitoring and evaluation**: Keep monitoring the output of your model. If you notice recurring mistakes of the model which could be bothersome for

[2]See for more information: https://christophm.github.io/interpretable-ml-book/shap.html

deployment domain workers, have discussions with them about how to alter the model. In the Reporting Issues case, this was e.g. done by adding new categories of public space nuisance to the model.
- **Organisational**: Assign specific roles within the organisation who are responsible for the connection between the model developers and the daily operations, for example a 'category manager' in the Reporting Issues case who makes sure that the model assigns citizen input in useful categories for the workers who resolve the issues.

6.4 Requirements Dilemma III: Responsive to Citizens Subjective Inputs and Fostering Objectivity

This section presents a dilemma of governance requirements which is based on some of the findings from the Reporting issues in public space case.

Observation 1: governments using AI should be responsive to citizen input

In the Reporting issues in public space case, citizen feedback on both the previous system and the first versions of the new system was used to form the new AIS of issue reporting for Amsterdam's public space. In this way, Amsterdam's citizens were able to contribute to the forming of the system. In itself, the online issue reporting system is a way in which citizens can shape their public space by partly determining where the CoA's public space resolving capacity is used for. Governments should seek to involve citizens with their decision-making and hence AIS development and, like in this case, AI sometimes offers possibilities to increase citizen participation with governmental decision-making. In this case, the participation also aided to gain trust with citizens:

"There was a question, if we as a municipality want to make more use of algorithms, how can we ensure that people keep their trust in us?" (...) Because if the image is created that we use algorithms in stead of people for our decision-making, who controls the algorithms?" - Interviewee I1

Observation 2: governments using AI should prevent new biases through citizen participation

In the same case, the possibility for citizens to provide feedback on the AIS also leads to new biases. The feedback is only provided by particular citizen subgroups, especially the ones with high-trust in government, outspoken citizens and higher educated citizens. If governments only listen to the select group of citizens that has input and comments, a new form of subjectivity and inconsistency might occur. To address this the CoA seeks for consistency of decision-making by not taking into account every single remark by citizens but stick to the data:

"We know we should not serve the more outspoken citizens better than the less outspoken ones, so we have to use the data as objectively as possible" - Interviewee I1

And if the AIS is completely based on citizen input, as is the case for Reporting Issues in Public Space, the CoA must not confuse the citizen input with the objective state of their city. Otherwise citizens who know how to find the CoA's systems, have benefits over others, according to this interviewee:

" It is not a great source for that. It's a nice indication, but reports can be filed anonymously. (...) Well, if you have a person who really likes to file reports about dogshit, then this will end up in the statistics that are used for decision-making. (...) It's an indication of course, but I would do it differently myself." - Interviewee I2

Governance requirements III: dilemma

The contradicting observations indicate a third governance requirements dilemma. If governments seek for citizen participation and transparency for their AISs, this leads to trust building between citizens and governments regarding their ADM use. As mentioned by Interviewee I1, this trust building is especially necessary in light of the current societal debates about governmental AI use. Governance requirement III-a therefore is:

Requirement III-a: **Be transparent about the AI systems you use. Communicate actively and provide opportunities for citizens' idea contribution and participation.**

On the other hand, the search for consistency, objective representations of reality and no preferential treatments for outspoken citizens result in this requirement:

Requirement III-b: **Foster objective representations of reality by your AI systems and prevent new forms of bias caused by citizen participation.**

AI System vulnerabilities leading to requirements dilemma III

The vulnerabilities model presents issues that emerge in the societal context before implementation, which concern the involvement of citizens with governmental AISs use. The following vulnerabilities, as well as the notion of current societal debate about ADM lead to governance requirement III-a:
Vulnerabilities - Societal context: "Unverifiable and faulty modeling thresholds", "Privacy infringements", and "Increasing information asymmetry between public agencies".

Governance requirement III-b relates to different benefits for citizens of public sector AI use which are presented in section 3.1.2:
Advantages: Benefits for citizens - "Potentially more objective decisions without partiality" and "Consistency of decisions". Requirement III-b is also connected to the following AIS vulnerability:
Vulnerabilities - Model deployment context: "New form of citizen dependence on unpredictable public sector organisations".

Mitigation measures for requirements III-a and III-b

Again, several mitigation measures found during the different case studies aid to meticulously fulfill governance requirement III-a:

- **Technological**: Investigate whether the model could optimise for citizen satisfaction parameters.
- **Monitoring and evaluation**: Keep track of citizen satisfaction with the system. If external parties are involved, make this an important KPI in the contract.

- **Organisational**: Invest in sufficient capacity to answer citizen questions and receive feedback on the system.
- **User engagement and citizen agency**: Be transparent about your AI systems. Examples are: create an openly accessible 'algorithm register' to inform citizens about the ADM system used by your government. Organise interviews with newspapers about your new algorithmic system. Organise presentations and information meetings with selected citizens.
- **User engagement and citizen agency**: If decisions are made or informed by ADM, inform citizens about this fact. For example when you communicate about the decision itself in a letter, or by clearly marking the scanning cars with "Parking enforcement".
- **User engagement and citizen agency**: Rather than only have them act in their watchdog role, actively involve societal actors and interest groups like Bits of Freedom and the *Ombudsman*.

The same holds for governance requirement III-b:

- **Technological**: Make sure your model is puristic. Do not let every specific claim from deploying department's employees or citizens influence your model development. For example in the Reporting issues in public space case, let the text classifying model stick to the input data and do not give priority to certain nuisance categories if colleagues ask to do so.
- **Organisational**: Discuss the findings from output data with political-administrative actors, to come up with long-term strategies to deal with emerging challenges.
- **Monitoring and evaluation**: Keep track of the objective state of your deployment domain in different ways besides only using the data your model generates. So e.g. in the Automated parking control case, do not only use the frequencies of data generated with the scanning cars but also send human employees to the city for observations about frequencies of parked cars.

6.5 Requirements Dilemma IV: Innovation and Upfront Goal Determination

This section presents a last governance requirements dilemma.

Observation 1: technology push and innovation bring merits

In the Reporting issues in public space case, the system SIA is, for the citizen input part, based on the automated ML model for automated text classification of citizen reports. That text classifier was an innovative project of one single development intern within the CoA, who presented it as an idea to the political-administrative actors who encountered problems with the old system. The idea was received with cheering and now the system functions to the satisfaction of both CoA workers and citizens. This innovative project, a technology push from one single developer, successfully set directions for the new decision-making system.

Observation 2: development teams should think about the goals of their projects

In the Automated parking control case, interviewees discussed the idea to add extra cameras to the scanning cars, besides the cameras for licence plate scanning, which could provide information about the state of Amsterdam's public space, e.g. potentially broken streetlights. However, it was not yet entirely clear what the goal of this extra scanning should be. In this case it was very questionable if this new technological solution actually solved an existing problem and if the involved actors agreed on

the question whether the merits of deploying new cameras would balance with the downsides of extra surveillance. The following quote reflects the opinion of one of the interviewees, who thought that many AI projects were too expensive and the contribution to goals of either the organisation or citizens remained unclear, leading the interviewee to think that other non-AI solutions would have been more optimal:

"We put all kinds of scrum masters on it en then we create more of these projects. Well, good luck. It's a nice idea, but it's too expensive and I notice in practice it can be done more quickly, more efficient and cheaper." - Interviewee I6

Governance requirements IV: dilemma

The abovementioned notions lead to a last governance requirements dilemma. First, governments have to be innovative and not fall behind on current developments in business and society. The importance of innovation is acknowledged by the European Commission (2021) as well: they call for 'regulatory sandboxes' to stimulate innovation in AISs. Technology push of innovative AI projects potentially lead to successful future solutions for challenges governments face, as seen in the Reporting issues in public space case. Governance requirement IV therefore is:

Requirement IV-a:
Stimulate innovation with regards to AI development within your organisation and do not restrict every innovative project upfront.

At the same time, innovation cannot always be a carte blanche and governments should be careful whether their AISs actually contribute to a public goal, and in some sensitive policy areas like fraud detection and law enforcement it should be even more careful for proportionality reasons. Governance requirement IV-b hence is:

Requirement IV-b:
Ensure that AI development projects have a clear and proportional contribution to an agreed policy objective.

AI system vulnerabilities leading to requirements dilemma IV

This governance requirements dilemma IV results from benefits of using AI by governments, as well as some AIS vulnerabilities. Governance requirement IV-a essentially stems from all benefits of public sector AI mentioned in section 3.1: if governments like the CoA want to reap the benefits of AI now and in the future, it needs to keep up with current developments and hence innovate.

Governance requirement IV-b relates to several vulnerabilities of AISs seen from the political-administrative context, before a model is implemented. These vulnerabilities all concern the lacking clarity of how an AI solution contributes to either governmental goals or citizens' well-being:
Vulnerabilities - Political-administrative context: "Discord with existing legislation and concerns about proportionality", "Relying on AI models undermines broader political debates", and "Weak relationship between AI deployment and public policy goals".

Mitigation measures for requirements IV-a and IV-b

This section presents the selected mitigation measures found in the case studies to fulfill requirements IV-a and IV-b. To start with governance requirement IV-a:

- **Organisational**: Set innovation through AI development as a goal in itself. In this way, innovation is safeguarded while also having a formal goal for projects.
- **Organisational**: Make sure there is a 'product owner' for every AIS development process. This is a non-developer who is responsible for the alignment and coordination between development teams and the political-administrative and model deployment actors. This business owner must ensure understanding between technological developers and the principals, in order to save time and resources from developers and create space for innovation.
- **Data**: Support developers by providing them with enough data to develop models. If necessary, this data can be manipulated to protect citizen privacy.

And for governance requirement IV-b:

- **Organisational**: Make sure there are formal agreements from the Board of Mayor and Aldermen about the necessity and use of the model. If that is done, make sure the model is actually adopted and used in the way agreed upon.
- **Organisational**: Actively engage the city council to inform them about the potential benefits and risks of using the specific AI solution.
- **Legal**: Involve legal experts to assess the proportionality of the model: how do the impact of the model on individuals and the impact of the model on societal or political goals compare with each other?

6.6 Chapter conclusion

This chapter presents several dilemmas of AI governance requirements found in the case studies, which follow from trying to steer the organisation in such a way that vulnerabilities are overcome. Mitigation measures from different categories to adhere to these requirements also became apparent in the case studies. These answer the sub-question for this chapter:

What are relevant governance requirements and mitigation measures, found during conduction of the case studies, to deal with context-dependent vulnerabilities of public sector AI systems?

Requirements dilemma I
The first set of requirements concerns the impact of the automated decisions. Governments want to increase impact of their models to overcome political-administrative vulnerabilities of weak relationships between AI and their policy goals, so increase the impact of the AI model on their decision-making process. At the same time, they seek to decrease the impact of models as to overcome AI model vulnerabilities regarding accuracy of decisions and biases.

Requirement dilemma II
This second set of requirements points towards a tension between commitments from the developers to create alignment with the organisational workflows, but not to tie these developers to a too large extent. Governments have to make sure the developers deliver a product which is accepted and understood within the application domain departments that will use it, to overcome the vulnerabilities of unaccepted or misunderstood products in the application domain. But at the same time, these same developers should be given enough trust and space. In that way they are able to

deliver high-end technological products which can actually live up to the promise of AI and overcome technological vulnerabilities as much as possible.

Requirement dilemma III
These requirements concern the citizen participation for public sector AISs. Seen from the growing information asymmetry and societal debates inflamed by ADM use, governments should put enough effort into the involvement of citizens with their algorithmic systems. At the same time, governments strive for objectivity and want to overcome the vulnerability of new forms of citizen dependency on governmental unpredictable AISs. This means they should take care not to adapt their tools to the outspoken and involved citizens only. Furthermore they sometimes have to simply make a choice as not all opinions and demands can be combined into a useful tool.

Requirements dilemma IV
This requirements dilemma concerns the potential tension between an innovative mindset and thinking about the contribution to policy goals of a new AI project upfront. Based on all potential benefits public sector AI brings, the CoA wants to keep up with developments and sometimes have projects just for innovative purposes. At the same time, based on the political-administrative contextual vulnerability of discord with legislation, undermining broader political debates and weak relationship between AISs and policy goals, the CoA also seeks for development teams that have a clear contribution in mind for their projects.

Chapter 7

Discussion

This discussion chapter reflects on the study results and their relevance, generically interprets the use of AI by governments and provides policy recommendations for governments using AI.

7.1 Reflection on the Research Results

In the literature reviewed for this thesis, many authors focus on the potential risks of ADM systems like AI. Oftentimes they look at the topic through the lens of their own, singular research discipline. One of those lenses for example is a computer science lens, when Corbett-Davies & Goel (2018) reflect on the mathematical impossibility of combining different algorithmic fairness metrics. Or an ethical lense, which helps Floridi et al. (2018) to come up with a framework for a 'Good AI Society'.

However, the challenges that AI adoption brings for governments are not bounded by scientific disciplines. If we for example look at the vulnerabilities that emerge in the model deployment context found in this study, there are, amongst others, technological challenges ("Incorrect input-output relationship"), public administration challenges ("Removing humans from decision-making can be a compromise on quality") and socioeconomic challenges ("Negative impact on workforce"). In the search for context-specific theory development and governance leads to deal with public sector AI risks, scientists should bring multiple research disciplines together to do justice to the complexity and interdependencies of the challenges governments face. This research has intended to do so by taking the different involved actors from the municipal practice as the groundwork for a theoretical framework, and then use all sources and disciplines deemed relevant for the vulnerabilities in different contexts.

The case study is conducted in the CoA, a large municipality. The question therefore arises whether the results from the study are generalisable to other kinds of governmental organisations. The main result of this study is chapter 4's vulnerabilities model. The case-specific vulnerabilities found in the case study actually serve to prove the generic character and validity of the model. The proven generic validity of the context-dependent vulnerabilities is a first argument for the model to be useful outside the scope of larger municipalities like Amsterdam. Secondly, the AI model context, political-administrative context and societal context are applicable to every governmental organisation doubtlessly. The model deployment domains will differ strongly between different kinds of governments, but the AI model is always intended to serve some department within the organisation, so this context is present in every

government as well. Lastly, the model has an inherent logic (see section 4.1.1) which also is an argument for the model to be usable outside Amsterdam's borders.

The AIS vulnerabilities found during the interviews showed that the lion's share of case-specific vulnerabilities are interpretable using the context-dependent generic vulnerabilities of chapter 4's model. This shows that the model is a useful tool for governments that want to consider the vulnerabilities of their AISs, either before or after implementing it. Nevertheless, the case study validation of the model also revealed three vulnerability-complements for the model: "Model being adopted without enough capacity to control it", "Inadequacies in the requisite security of the AI model", and 'Unwanted strategic behaviour with the AI model". These complements are useful lessons in itself, but also prove that studying specific cases with the vulnerabilities model will probably always lead to new insights of vulnerabilities besides the generic ones already presented in the model. This implicates that the model is useful to think about vulnerabilities from the different context perspectives, and also to have a first overview of often encountered vulnerabilities within these contexts, but it should definitely not be used as an extensive or exhaustive list of all possible risks and vulnerabilities in practice.

Chapter 6 reports about what the case studies teach us about how vulnerabilities of AISs result in governance requirements for public sector decision-makers. Interestingly, the results show how trying to manage vulnerabilities emerging in one context, often lead to tension with other governance requirements. Increasing impact of an AIS on the governmental operational chain to make sure it serves a policy goal comes at the cost of increased impact for citizens in case of a model mistake as well, for example. The results therefore do not only aid decision-makers by increasing their understanding of what can go wrong, but also further problematise their governance strategy by revealing trade-offs and dilemmas. This finding is in line with previous literature findings which indicate that AI governance often leads to trade-offs and hard choices. Further problematisation of the governance challenges also fits the needs from governmental practice. Mitigation measures like privacy assessments, technological optimisation and monitoring are well-known and available. The challenge lies in understanding the complexity of the challenges using ADM systems bring and what are the key decisions to make. The vulnerabilities model and governance requirements are of help here.

The Automated parking control case and Illegal housing rental case both concern ADM systems focused on enforcement. The AI techniques only play a minor role in this enforcement: to recognise license plates automatically and determine the order in which enforcement officers check complaints respectively. Using algorithmic systems is inescapable, for both governments and business. Discussions about algorithms in this light must concern how and for what reasons governments use them. When there is a discussion about fairness, potential discrimination or bias, and other vulnerabilities of the AIS, this discussion must inevitably be broader: how does a governmental organisation want to relate with its citizens? Regarding the harshness of punishments based on automated decisions, the possibility of model mistakes compared with the generic interests it serves, and extensiveness of controlling and checking citizens: what is acceptable? By focusing on the vulnerabilities of AISs in this research, that discussion might even be compromised. That is, because drawing up vulnerabilities of AISs might create the false idea that these are just challenges that could be overcome, leading to safe use of AI. The Reporting issues in public space points towards different kinds of discussions, as the starting point for this AIS is not to check and potentially penalise citizens. Here the main issue is to deal with the AIS vulnerabilities rather than having an upfront discussion about the role of a government.

7.1.1 Scientific Relevance of the Results

Current literature which focuses on ethical aspects of AI, seems to reach agreement on a shared concern. That is, the downsides of deploying AI in the public sector have been undervalued or sometimes even neglected in the early days of its emergence. Recently, authors began to focus on downsides for citizens like bias, discrimination, lack of transparency and others. Many of the articles which shed governance-focused light on AI are top-down and essayist. Ziewitz (2016) for example writes a governance-focused essay about the 'myth of algorithms'. He states that the goal cannot be to come up with detailed instructions, but rather should be to stay generative when considering algorithms. On the contrary, this thesis contributes by using a bottom-up approach based on the combination of a conceptual framework and lessons learned from the CoA's daily practice.

The scientific relevance is a logical consequence of taking steps in addressing the scientific knowledge gaps as presented in section 1.3. Summarising, there is a need for concise knowledge about citizen harms originating from public use of AISs, and what governance trade-offs occur when dealing with them. An aspect which contributes to the scientific importance is to take an integrative approach by approaching the topic from a system perspective, as described in section 1.2.2. This integrative approach aids to find out whether AIS problems which do harm to citizens can actually be dealt with in a way which allows for social acceptance (Wirtz et al., 2019). Scholars must approach citizen safety issues of public AISs as sociotechnical phenomena and only strong synergy between social and technological insights can lead to safe public sector AI (Dobbe & Raji, 2019). This thesis takes such an approach.

7.1.2 Societal Relevance of the Results

Like all local governments, the CoA is involved in decisions which shape the public space, employment opportunities, inclusiveness of cities and so forth (GroenLinks, D66, PvdA, & SP, 2018). This means their AISs play an increasing role in very high-stake domains for citizens. Growing media attention reflects this fact, illustrated by for example this Dutch newspaper interview about an AIS for tax policy reform (Witteman, 2020). And, as presented in the research context in section 1.1, AI can cause severe consequences for citizens in The Netherlands as well. This thesis contributes to understanding and potentially better mitigation of these vulnerabilities.

This thesis intends to profile public AI vulnerabilities into a workable construct which lays the groundwork for better governance. This is a first step in overcoming the gap between bottom-up development by technological experts and e.g. Chief Information Officers who seek to understand and minimise risks and negative impacts of the algorithmic systems their public organisations uses. Public managers now mostly use the Dutch privacy law *Algemene Verordening Gegevensbescherming* and the information security baseline *Baseline Informatiebeveiliging Overheid (BIO)* for steering and quality checks of algorithms. However, the public sector seeks for relevant frameworks and directives which are specific for ADM and do justice to the broad societal discussions about algorithms (Algemene Rekenkamer, 2021).

7.2 Generic Reflection on the Public Sector Using AI

Governments like the CoA see the benefits AI potentially brings to its organisation or the citizens in Amsterdam. Furthermore, they are a part of society, in which AI plays an increasingly prominent role, which is another reason to keep up with these developments. Citizens expect their governments to do so as well. But this thesis

shows a very complex situation of context-specific and at the same time interdependent vulnerabilities of public sector AISs. These may ultimately result in harms to citizens. The three cases studied also show that diving into specific AISs for different deployment contexts will inevitably lead to new insights on more vulnerabilities. This might seem overwhelmingly complex. The question then is: what is the overall challenge and how do first steps towards resolving look like?

The overall challenge is not to prevent, mitigate or compensate every vulnerability of public sector AISs, which would be a Utopian mission. Because using techniques like AI for ADM is a relatively new development, the challenge is to increase the maturity of organisations with regards to governance strategies for AISs:

i) Governments acknowledge the urgency to protect citizens from AI-induced harms[1], but lack knowledge of the variety of interdependent vulnerabilities. Development teams within governments like the CoA lack oversight over the assembly of risks that AI might pose, and how these risks are specific to e.g. the involved departments, context of deployment, technological difficulties, and impact of the ADM on citizens' well-being. Such challenges are of technological, ethical, socioeconomic, democratic, and legal nature.
ii) Only when development teams are aware of the variety of vulnerabilities, the responsible decision-makers are enabled to make the detailed and structured considerations required for AISs that have significant impacts on citizen well-being. If such considerations are made by the department managers, who carry responsibility to the city council and Board of Mayor and Aldermen, or the city council and Board itself, the important choices become a matter of societal and political debate, which is what the high-impact AISs ask for.

The interesting considerations to be made are threefold. First, responsible decision-makers could argue for vulnerabilities not to play a role in their specific case. Secondly, if vulnerabilities do play a role, they can make a case from the available risk control measures. Such mitigation measures are categorised in section 3.3 and vary from Privacy Impact Assessments to model developers choosing explainability of model architecture over performance indicators like accuracy. Thirdly, for the vulnerabilities present and not or incompletely mitigated, decision-makers must choose whether they find the potential impact of vulnerabilities acceptable, or prefer the mitigation of one risk over the other, or to consider the advantages the system brings to outweigh its vulnerabilities.

Concluding, it is not the existence of AIS vulnerabilities, but merely the absence of a thorough process to settle considerations about the vulnerabilities which is the challenge for governments now. If governments use reports like this to understand the context-specific and interdependent vulnerabilities, this forms a first step towards this process and creating room for politically responsible decision-makers to make the relevant trade-offs and in this way allow societal and political debate about the appropriateness of their decisions. The next step is to document such considerations and create overviews of best practices, so that learning between development teams and their managers within or outside the CoA and other governmental bodies is allowed.

[1] *"We find ourselves in a situation, and I think that should be taken into account, in which there is a large societal discussion about the use of algorithms"* - Interviewee I7

7.3 Policy Recommendations

The CoA has essentially defined two goals for digital technologies in its policy papers. The first is to bring added value in model deployment domains for either the organisation or citizens, e.g. in (Amsterdam Municipality, 2019). The second is to use them for innovative purposes, which is defined as a goal in itself, e.g. in (Gemeente Amsterdam, 2018). One of the most common discussions during the last months of doing research at the CoA was about which of these policy goals is served by the AIS at hand, as this often remained ambiguous or simply undefined. As with every other decision-making strategy, AISs bring vulnerabilities. So to have a clear purpose in mind for the AIS that might be developed is the most important gatekeeper for AI projects. This purpose should be a shared agreement of at least actors from the model context, model deployment context, and political-administrative context. Preferably actors from the societal context are involved with these discussions as well, they can be represented by selected diverse groups of citizens or citizen interest groups. Goals for the AIS can develop over time, but to at least have a shared belief in what the system is ought to do and keep discussing this over the lifecycle of the system, helps to find out about the relevant consideration of vulnerabilities as just described.

Besides having an agreed purpose for the AIS, the distinction between efficiency goals, effectiveness goals, citizen well-being goals and innovative goals should always be made clearly as well. This manages expectations and aids in honest and clear communication with citizens. In case the goal is to bring about an impact in the deployment domain, the impact of the model on the decision-making chain should be large enough to at least start the discussion whether the vulnerabilities it brings are acceptable. This is reflected in governance requirement I-a from chapter 6: "Increase the impact of the AI model on the decision-making process to ensure its added value and create balance with the downsides of the AI system". If innovation is the goal of governmental AI development, then the innovation must still have a purpose, for example expected use cases in other domains or later in time. Innovation should never be a cover-up for pointless digital projects. Technological opportunities are not to direct decision-makers, but rather the question what is the role of your government for citizens and how technological opportunities can be of help in fulfilling that role.

The governance requirements based on context-dependent AIS further problematise using AI in the public sector. If you leave space for developers to construct high-end UL algorithms which overcome technological vulnerabilities before implementation, you might compromise model understanding with deploying department's workers after implementation. Resolving challenges in one context, often seems to lead to emergence of vulnerabilities in other contexts. Understanding such tensions and trade-offs should be the primary focus when assessing the risks of using ADM within governments. Using the vulnerabilities model can be of help to do so: multidisciplinary teams with actors from all contexts can use the model for directions to think about the vulnerabilities they encounter. It is strongly recommended to therefore always engage actors from all contexts with the development teams, as seen for the model context, model deployment context and political-administrative context in the "Werkplaats" for Illegal housing rental.

Teams often do not engage societal context actors, due to for example perceived lack of interest or understanding from citizens. This is an invalid argument, as in such cases there are always societal actors like interest groups who could be involved. Some CoA employees mistakenly interpret citizen participation as a step towards the inclusion of all citizens' interests into the AIS. But that is neither possible nor the purpose of citizen participation. In light of this research, including citizens in the development process will lead to a better understanding of vulnerabilities within the societal context,

by incorporating their perspective on advantages and disadvantages the system brings. A positive example of such participation is found in the Reporting issues in public space case, where citizen feedback on both the previous system and plans for a new system played a role in understanding the needs for the new system. Or the Illegal housing rental case, where a group called "Commissie Persoonsgegevens Amsterdam" of journalists, lawyers, philosophers and others de facto acted as a delegation for the societal context. Transparency measures like publishing algorithmic systems in the "Algoritmeregister" are therefore not just a showpiece. Although only being transparent is not enough to deal with AIS vulnerabilities, it is essential and thus highly recommendable to provide societal actors with more opportunities to get involved with governmental use of AI. Media, interpreted as societal actors, are then enabled to take up their role as well.

In all studied cases, interviewees discussed how their teams tried to deal with the potential biases and unfairness caused by the algorithmic system. But decision-makers should not lose oversight by primarily focusing on the bias of the algorithms themselves. If algorithms determine the chance of Airbnb rental fraud based on personal information, there will be bias towards some household sizes, house sizes, and area codes. If algorithmic systems facilitate reports about public space as a means for public space management, there will be bias towards less outspoken citizens. If external scanning car providers have financial incentives to scan as many cars as possible, there will be bias towards car-dense areas in Amsterdam. Bottom line is that decision-makers must analyse the bias that occurs *by using the AI* rather than the *bias of the AI itself*. This is oftentimes not so much the problem, as in the cases Reporting issues in public space and Automated parking control, or it is the exact reason for the algorithmic system to be used in the first place, as in the Illegal housing rental case. Only focusing on mitigation of the algorithmic bias itself would therefore miss the point.

Chapter 8

Conclusion

This chapter answers the main research question, presents the research limitations, and sets directions for further research.

8.1 Answer to the Main Research Question

This thesis aims to answer the following research question:

In public sector AI systems, what are emerging vulnerabilities for citizens and how do these translate into governance requirements for decision-makers?

The "emerging vulnerabilities for citizens" part of the main research question concerns the **context-dependent model of vulnerabilities for citizens**. Context-dependency in AI governance is often interpreted as dependency with regards to the field of AI application, for example vulnerabilities specific to enforcement AISs or public space AISs. During the preliminary discussions with CoA employees and the literature review, another context-dependency of vulnerabilities came forward. AIS vulnerabilities strongly differ between perspectives of the involved departments and societal actors in the governmental practice. When governments decide to use AI to make or support decisions, different groups get involved. First there are the model developers, represented by the AI model context. There is the department for which an AI solution is constructed, the model deployment context. The last governmental context is the political-administrative, where governmental managers and politicians carry final responsibility for the well-functioning of the model deploying departments. The fourth and last context is the societal context, formed by the citizens the governmental organisation essentially seeks to serve.

Although it is cumbersome to repeat all vulnerabilities in this conclusion, an interpretation of the common vulnerabilities helps to illustrate the context-dependency of vulnerabilities. Vulnerabilities in the AI model context mainly are technological and data vulnerabilities, as well as lacking application domain understanding with model developers. Such vulnerabilities could lead to citizen harms like wrong court convictions due to low-accuracy DNA matching AI technologies. In the model deployment context, vulnerabilities include problems of representing the real-world situation with quantified methods like AI, and negative impacts on existing governmental workforce. Citizens harms are caused if AI models are not equipped to adapt to new demands in its application reality. Vulnerabilities in the political-administrative context are ambiguity about how AI contributes to reaching public goals and legal concerns about for example the presumption of innocence being compromised when governments use

predictive algorithms for enforcement. This leads to citizen harms when people get accused of social benefits fraud based on personal information which is obtained in an illegal way. Lastly, telling vulnerabilities in the societal context are diminishing possibilities to control governments for citizens and contribution to existing societal disparities. Citizen harms occur when for example the Dutch government only uses its automated fraud detection system SyRI in poorer urban areas.

Lastly, **using the vulnerabilities model, governance requirements found in the case study** are presented to answer the main research question. Vulnerabilities lead to governance requirements, but these requirements among themselves lead to trade-offs which governments like the CoA have to address. So, to use requirements set I to illustrate: if governments want to overcome the vulnerability from political-administrative perspective that AISs do not contribute to policy goals, impact of the model on decision-making has to be increased, for example by completely replacing human capacity. At the same time, AI model technological vulnerabilities require governments to neutralise the impact of models on their decision-making, for example by adding human capacity to check model decisions. Governments have to be aware of the trade-offs between governance requirements which follow from the mitigation of vulnerabilities in different contexts.

8.2 Research Limitations

A first limitation of this study is the lacking expert validation of the contextual vulnerabilities model. Priority for the case studies was to interview involved CoA employees with different expertise. Due to time constraints for a master's thesis, the choice is made not to invest in interviews with experts from inside and outside the CoA to discuss the constructed theory with. Such discussions would have contributed to sharpening the ideas about context-dependent vulnerabilities themselves, as well as the relationships between the four different contexts. However, the case study interviews themselves proved a useful validation method as well. The lion's share of the case-specific vulnerabilities of Amsterdam's AISs are interpretable using the context-dependent vulnerabilities of chapter 4's theoretical model. This shows that this generic framework encompasses a large part of the vulnerabilities that governments encounter in practice.

All discussions about the vulnerabilities of AISs which may lead to harms for citizens were held with a wide range of involved actors, except for citizens themselves. The reason to invite CoA workers from the deploying departments, developers and managers was to get a nuanced view on the vulnerabilities which emerge in different contexts, as well as to find directions for governance requirements and mitigation measures from different angles. To interview citizens from Amsterdam would have required a different interview approach. Rather than asking for challenges and best practices the interviewees encounter in their work on the AIS, these interviews would have concerned general concerns with governmental use of AI. Interviewing citizens would furthermore have demanded to carefully select a balanced group of citizens, with a prerequisite that every interviewee had some basal knowledge of ADM. Because of the focus on theory development and finding governance requirements, combined with the accessibility of interviewees from the CoA and limited time resources, a focus on experts who encounter AI in their daily practice is chosen. To not involve citizens is somewhat paternalistic and comes at cost of understanding vulnerabilities in the societal context.

The third limitation concerns the selected cases. Due to the availability of preparatory information through the municipality's Algorithm Register (City of Amsterdam, n.d.),

as well as the established connections with the CoA supervisor and people involved with these cases, three cases which are already represented in the register are selected. The fact that these cases are incorporated in the register, does not mean that the risk management of the algorithmic systems is anywhere near perfect, but it does mean that there is a lower bound of how the involved teams deal with the downsides of the system. Otherwise the municipality would probably not have published these systems in the online and openly accessible register. To have had the time to also find cases in which the current dealing with AI risks was of much lower quality, more chaotic or maybe even completely absent, would have been interesting for the theory development and governance requirements of this thesis. Presumably, more complements of the vulnerabilities model within the model deployment context and political-administrative context would have been found whilst studying such chaotic cases.

8.3 Future Research

Several directives for future research would contribute to the findings from this thesis.

First of all, the case study of this thesis only looked into three cases. Involving more different cases would contribute by providing more ideas for complementing the vulnerabilities model, as well as lead to more governance requirements based on the vulnerabilities. It would be especially interesting to involve a case where predictive algorithmic systems are used for citizen support rather than enforcement. An example is the use of AI to predict which citizens are likely to fall into debt in the near future, so that governments could offer support programs to them. Involving such cases will most likely direct to other sorts of vulnerabilities and biases. Model mistakes are of less importance because these do not lead to citizen harms. In such cases, harmful effects occur if the model misses out on citizens in need of support.

Another suggestion is to involve citizens in the interviews for case studies. Involving actors from all four contexts of the vulnerabilities model rather than only the actors which were accessible through the CoA municipal channels will help to nuance the vulnerabilities found in the societal context as well as lead to new insights in governance requirements regarding citizen participation. Citizens most likely also will come up with out-of-the-box mitigation measures, because their ideas are not formed by pre-existing knowledge and practices.

Lastly it is recommendable for other researchers to actually get involved with governmental AI development teams. For researchers with a governance lens to actually join such teams up from the beginning of an AI development trajectory will give useful insights in the dynamics between actors from the different contexts of this research: how do political actors engage with the developers for example? This research already tried to grasp these dynamics by interviewing the involved CoA employees and explicitly ask questions about the feedback between developers, deploying department workers, political-administrative actors and citizens. However the lessons learned from these dynamics are more valuable if they i) come from first hand and ii) are learned throughout the whole development cycle of a specific AIS.

References

AI HLEG. (2019). *Ethics Guidelines for Trustworthy AI* (Tech. Rep.). Brussels: European Commission. Retrieved from https://ec.europa.eu/futurium/en/ai-alliance-consultation

AI HLEG. (2020). *The Assessment List For Trustworthy Artificial Intelligence (ALTAI): For Self Assessment* (Tech. Rep.). Brussels: European Commission. Retrieved from https://ec.europa.eu/digital-single-market/en/news/assessment-list-trustworthy-artificial-intelligence-altai-self-assessment

AI Now Institute. (2018). *Litigating Algorithms: Challenging Government Use of Algorithmic Decision Systems* (Tech. Rep.). New York City: AI Now Institute. Retrieved from https://ainowinstitute.org/litigatingalgorithms.pdf

Algemene Rekenkamer. (2021). *Aandacht voor Algoritmes* (Tech. Rep.). The Hague: Algemene Rekenkamer. Retrieved from http://www.surfsharekit.nl:8080/get/smpid:15947/DS1

Amsterdam Municipality. (2019). *A Digital City for and by Everyone* (Tech. Rep. Nos. Agenda for the Digital City, version 1). Amsterdam: Gemeente Amsterdam. Retrieved from https://www.amsterdam.nl/wonen-leefomgeving/innovatie/de-digitale-stad/

Androutsopoulou, A., Karacapilidis, N., Loukis, E., & Charalabidis, Y. (2019). Transforming the communication between citizens and government through AI-guided chatbots. *Government Information Quarterly*, *36*(2), 358–367. doi: 10.1016/j.giq.2018.10.001

Auerbach, C. F., & Silverstein, L. B. (2003). *Qualitative Data: An Introduction to Coding and Analysis* (First ed.). New York: New York University Press. Retrieved from https://books.google.nl/books/about/Qualitative{_}Data.html?id=6u{_}FPbXSmBQC{&}redir{_}esc=y

Autoriteit Persoonsgegevens. (2020). *Belastingdienst/Toeslagen: De Verwerking van de Nationaliteit van Aanvragers van Kinderopvangtoeslag* (Tech. Rep.). Den Haag: Autoriteit Persoonsgegevens. Retrieved from https://autoriteitpersoonsgegevens.nl/sites/default/files/atoms/files/onderzoek{_}belastingdienst{_}kinderopvangtoeslag.pdf

Barocas, S., & Selbst, A. D. (2016). Big Data ' S Disparate Impact. *California Law Review*, *104*, 671–732.

Bogner, A., Littig, B., & Menz, W. (2009). Introduction: Expert Interviews - An Introduction to a New Methodological Debate. In A. Bogner, B. Littig, & W. Menz (Eds.), *Research methods series: Interviewing experts* (First ed., pp. 1–13). Hampshire, United Kingdom: Palgrave MacMillan. Retrieved from https://link.springer.com/book/10.1057/9780230244276

Bogner, A., & Menz, W. (2009). The Theory-Generating Expert Interview: Epistemological Interest, Forms of Knowledge, Interaction. In A. Bogner, B. Littig, & W. Menz (Eds.), *Research methods series: Interviewing experts* (First ed., pp.

43–80). Hampshire, United Kingdom: Palgrave MacMillan. Retrieved from https://link.springer.com/book/10.1057/9780230244276

Boyd, D., & Crawford, K. (2012). Critical Questions for Big Data: Provocations for a Cultural, Technological, and Scholarly Phenomenon. *Information Communication and Society*, *15*(5), 662–679.

Briscoe, E., & Feldman, J. (2011). Conceptual complexity and the bias / variance tradeoff. *Cognition*, *118*(1), 2–16. doi: 10.1016/j.cognition.2010.10.004

Chowdhury, R., & Sloane, M. (2020). *The Risks of Using AI for Government Work*. Retrieved 2021-01-26, from https://www.brinknews.com/the-risks-of-using-ai-for-government-work/

Chun, A. H. W. (2007). Using AI for e-Government automatic assessment of immigration application forms. In *Proceedings of the national conference on artificial intelligence* (Vol. 2, pp. 1684–1691). Retrieved from https://www.aaai.org/Papers/AAAI/2007/AAAI07-273.pdf

Citron, D. K., & Pasquale, F. (2014). The scored society: Due process for automated predictions. *Washington Law Review*, *89*(1), 1–33.

City of Amsterdam. (n.d.). *Algorithm Register*. Retrieved 2021-01-06, from https://algoritmeregister.amsterdam.nl/en/ai-register/

Cohen, I. G., Amarasingham, R., Shah, A., Xie, B., & Lo, B. (2014). The legal and ethical concerns that arise from using complex predictive analytics in health care. *Health Affairs*, *33*(7), 1139–1147.

Corbett-Davies, S., & Goel, S. (2018). The measure and mismeasure of fairness: A critical review of fair machine learning. *arXiv*, 1–25.

Danaher, J. (2016). The Threat of Algocracy: Reality, Resistance and Accommodation. *Philosophy and Technology*, *29*(3), 245–268.

Danaher, J., Hogan, M. J., Noone, C., Kennedy, R., Behan, A., De Paor, A., ... Shankar, K. (2017). Algorithmic governance: Developing a research agenda through the power of collective intelligence. *Big Data and Society*, *4*(2), 1–21. doi: 10.1177/2053951717726554

de Bruijn, H., & Herder, P. M. (2009). System and actor perspectives on sociotechnical systems. *IEEE Transactions on Systems, Man, and Cybernetics Part A:Systems and Humans*, *39*(5), 981–992. doi: 10.1109/TSMCA.2009.2025452

de Ruijter, M. (2021). *Parkeerboetes door scanauto ' s aanvechten loont*. Retrieved from https://nos.nl/artikel/2363617-parkeerboetes-door-scanauto-s-aanvechten-loont.html

Dijkstra, G. (2021). *Deze zoektocht naar 'boodschappenfraude' is ondemocratisch*. Retrieved from https://www.trouw.nl/opinie/deze-zoektocht-naar-boodschappenfraude-is-ondemocratisch{~}b40754b7/

Dobbe, R. (2020a). *TB242IA Intelligende Data-Analyse: Bias in Predictie, Aannames voor Predictie* (No. December). Delft: TU Delft.

Dobbe, R. (2020b). *TB242IA Intelligente Data-Analyse: Linear Regression Continued, Multiple Linear Regression* (No. November). Delft: TU Delft.

Dobbe, R., Dean, S., Gilbert, T., & Kohli, N. (2018). A Broader View on Bias in Automated Decision-Making: Reflecting on Epistemology and Dynamics. *arXiv preprint [Cs, Math, Stat] , arXiv:1807.00553*, 1–5.

Dobbe, R., Gilbert, T. K., & Mintz, Y. (2021). Hard Choices in Artificial Intelligence. *NeurIPS*, 242–242. doi: 10.1145/3375627.3375861

Dobbe, R., & Raji, I. D. (2019). Concrete Problems in AI Safety, Revisited. In *International conference on learning representations (iclr) 2020*.

Eggers, W. D., Fishman, T., & Kishnani, P. (2017). *AI-augmented human services: using cognitive technologies to transform program delivery* (Tech. Rep.). Deloitte Insights. Retrieved from https://www2.deloitte.com/content/dam/insights/us/articles/4152{_}AI-human-services/4152{_}AI-human-services.pdf

Eisenhardt, K. M. (1989). Building Theories from Case Study Research. *Academy of Management Review*, *14*(4), 532–550.

European Commission. (2021). *Proposal for a Regulation: Laying Down Harmonised Rules on Artificial Intelligence* (Tech. Rep.). Brussels: European Union. Retrieved from https://digital-strategy.ec.europa.eu/en/library/proposal-regulation-laying-down-harmonised-rules-artificial-intelligence

Fan, H., Bennetts, V. H., Schaffernicht, E., & Lilienthal, A. J. (2019). Towards Gas Discrimination and Mapping in Emergency Response Scenarios Using a Mobile Robot with an Electronic Nose. *Sensors (Switzerland)*, *19*(3). doi: 10.3390/s19030685

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., ... Vayena, E. (2018). AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations. *Minds and Machines*, *28*(4), 689–707. doi: 10.1007/s11023-018-9482-5

Flyvbjerg, B. (2006). Five misunderstandings about case-study research. *Qualitative Inquiry*, *12*(2), 219–245.

Friedman, B., & Nissenbaum, H. (1996). Bias in Computer Systems. *ACM Transactions on Information Systems (TOIS)*, *14*(3), 330–347. doi: 10.1145/230538.230561

Gemeente Amsterdam. (2018). *Informatievisie (i-visie)* (Tech. Rep.). Amsterdam: Gemeente Amsterdam. Retrieved from https://www.amsterdam.nl/bestuur-organisatie/volg-beleid/bestuur-organisatie/visie/

Gervais, D. (2020). Is Intellectual Property Law Ready for Artificial Intelligence? *GRUR International*, *69*(2), 117–118. doi: 10.1093/grurint/ikz025

Ghahramani, Z. (2015). Probabilistic machine learning and artificial intelligence. *Nature*, *521*(7553), 452–459. doi: 10.1038/nature14541

GroenLinks, D66, PvdA, & SP. (2018). *Coalitieakkoord Amsterdam: Een nieuwe lente en een nieuw geluid* (Tech. Rep.). Amsterdam: Gemeente Amsterdam.

Hadfield-Menell, D., Milli, S., Abbeel, P., Russell, S., & Dragan, A. D. (2017). Inverse reward design. In *Advances in neural information processing systems* (pp. 6766–6775). Long Beach, CA, USA.

Hildebrandt, M. (2015). Criminal Law and Technology in a Data-Driven Society. In M. D. Dubber & T. Hörnle (Eds.), *The oxford handbook of criminal law.* Oxford, United Kingdom: OUP Oxford. doi: 10.1093/oxfordhb/9780199673599.013.0009

Hill, R. K. (2016). What an Algorithm Is. *Philosophy and Technology*, *29*(1), 35–59. doi: 10.1007/s13347-014-0184-5

Hind, M., Wei, D., Campbell, M., Codella, N. C., Dhurandhar, A., Mojsilović, A., ... Varshney, K. R. (2019). TED: Teaching AI to explain its decisions. In *Proceedings of the 2019 aaai/acm conference on ai, ethics, and society* (pp. 123–129). Honolulu: IBM Research AI. Retrieved from https://dl.acm.org/doi/pdf/10.1145/3306618.3314273 doi: 10.1145/3306618.3314273

Huisman, C. (2020). *Fraudeopsporingssysteem SyRI schendt mensenrechten , overheid moet ermee stoppen.* Retrieved from https://www.volkskrant.nl/nieuws-achtergrond/fraudeopsporingssysteem-syri-schendt-mensenrechten-overheid-moet-ermee-stoppen{~}b83c21da/

Janssen, M., Hartog, M., Matheus, R., Yi Ding, A., & Kuk, G. (2020). Will Algorithms Blind People? The Effect of Explainable AI and Decision-Makers' Experience on AI-supported Decision-Making in Government. *Social Science Computer Review*, 1–16.

Janssen, M., & Kuk, G. (2016). The challenges and limits of big data algorithms in technocratic governance. *Government Information Quarterly*, *33*(3), 371–377.

Janssen, M., van der Voort, H., & Wahyudi, A. (2017). Factors influencing big data decision-making quality. *Journal of Business Research*, *70*, 338–345.

Johnson, M. G. (2020). City in Code: The Politics of Urban Modeling in the Age of Big Data. *Open Philosophy*, *3*(1), 429–445. doi: 10.1515/opphil-2020-0115

Kirchner, L. (2017, sep). *Thousands of Criminal Cases in New York Relied on Disputed DNA Testing Techniques*. New York City. Retrieved from https://www.propublica.org/article/thousands-of-criminal-cases-in-new-york-relied-on-disputed-dna-testing-techniques

Kitchin, R. (2014). *The Data Revolution: Big Data, Open Data, Data Infrastructures and Their Consequences* (R. Rojek, Ed.). London: SAGE Publications. doi: 10.29085/9781783302598.020

Kleinberg, J., Mullainathan, S., & Raghavan, M. (2017). *Inherent Trade-Offs in the Fair Determination of Risk Scores* (Vol. 67). doi: 10.4230/LIPIcs.ITCS.2017.43

Klievink, B., Romijn, B. J., Cunningham, S., & de Bruijn, H. (2017). Big data in the public sector: Uncertainties and readiness. *Information Systems Frontiers*, *19*(2), 267–283.

Kolkman, D. A., Campo, P., Balke-Visser, T., & Gilbert, N. (2016). How to build models for government: criteria driving model acceptance in policymaking. *Policy Sciences*, *49*(4), 489–504. doi: 10.1007/s11077-016-9250-4

Krafft, P. M., Young, M., Katell, M., Huang, K., & Bugingo, G. (2020). Defining AI in policy versus practice. In *Aies 2020 - proceedings of the aaai/acm conference on ai, ethics, and society* (pp. 72–78). New York. doi: 10.1145/3375627.3375835

Kroll, J. A., Huey, J., Barocas, S., Felten, E. W., Reidenberg, J. R., Robinson, D. G., & Yu, H. (2017). Accountable Algorithms. *University of Pennsylvania Law Review*, *165*, 633–705.

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

Lecun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444.

Linders, D. (2020). *Landmark ruling in SyRI case : Dutch court bans risk profiling*. Retrieved 2021-05-02, from https://solv.nl/blog/landslide-victory-in-syri-case-dutch-court-bans-risk-profiling/

Lison, P. (2012). *An introduction to machine learning*. Oslo: University of Oslo. Retrieved from https://www.nr.no/{~}plison/pdfs/talks/machinelearning.pdf

Lu, H., Li, Y., Chen, M., Kim, H., & Serikawa, S. (2018). Brain Intelligence: Go beyond Artificial Intelligence. *Mobile Networks and Applications*, *23*(2), 368–375. doi: 10.1007/s11036-017-0932-8

Mak, A. (2017). *Can You Trust Navigation Apps During a Major Emergency?* Retrieved 2021-03-30, from http://www.slate.com/blogs/future{_}tense/2017/12/07/california{_}wildfires{_}raise{_}questions{_}about{_}whether{_}you{_}can{_}trust{_}waze{_}.html

Mateescu, A., & Elish, M. C. (2019). *AI in Context: The Labor of Integrating New Technologies* (Tech. Rep.). New York: Data & Society Institute. Retrieved from https://datasociety.net/wp-content/uploads/2019/01/DataandSociety{_}AIinContext.pdf

Meuser, M., & Nagel, U. (2009). The Expert Interview and Changes in Knowledge Production. In A. Bogner, B. Littig, & W. Menz (Eds.), *Research methods series: Interviewing experts* (First ed., pp. 17–42). Hampshire, United Kingdom: Palgrave MacMillan. Retrieved from https://link.springer.com/book/10.1057/9780230244276

Milakovich, M. E. (2012). Anticipatory Government: Integrating Big Data for Smaller Government. *Internet, politics, policy 2012: Big data, big challenges.*, 1–13.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., ... Hassabis, D. (2015). Human-level control through deep reinforcement learning. *Nature*, *518*(7540), 529–533.

Morley, J., Elhalal, A., Garcia, F., Kinsey, L., Mökander, J., & Floridi, L. (2021). Ethics as a Service: a Pragmatic Operationalisation of AI Ethics. *arXiv*, 1–21.

Narayanan, A. (2019). *How to recognize AI snake oil*. Princeton University & Center for Information Technology Policy. Retrieved from https://www.cs.princeton.edu/{~}arvindn/talks/MIT-STS-AI-snakeoil.pdf

Olteanu, A., Castillo, C., Diaz, F., & Kıcıman, E. (2019). Social Data: Biases, Methodological Pitfalls, and Ethical Boundaries. *Frontiers in Big Data*, *2*(13). doi: 10.3389/fdata.2019.00013

O'Neill, C. (2016). *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy* (First ed.). New York: Crown Publishers. Retrieved from https://books.google.nl/books?hl=nl{&}lr={&}id=NgEwCwAAQBAJ{&}oi=fnd{&}pg=PA1{&}dq=weapons+of+math+destruction{&}ots=HxsjKoP3-d{&}sig=X8g8TKLO5tBpxHtRnlu8p-j1MVo{&}redir{_}esc=y{#}v=onepage{&}q=weaponsofmathdestruction{&}f=false

Parlementaire Ondervragingscommissie Kinderopvangtoeslag. (2020). *Ongekend Onrecht* (Tech. Rep.). Den Haag: Tweede Kamer der Staten-Generaal. Retrieved from https://www.tweedekamer.nl/sites/default/files/atoms/files/20201217{_}eindverslag{_}parlementaire{_}ondervragingscommissie{_}kinderopvangtoeslag.pdf

Pennachin, C., & Goertzel, B. (2007). *Artficial General Intelligence* (Second ed.; C. Pennachin & B. Goertzel, Eds.). Rockville: Springer. Retrieved from https://link.springer.com/content/pdf/10.1007/978-3-540-68677-4.pdf

Pfadenhauer, M. (2009). At Eye Level: The Expert Interview – a Talk between Expert and Quasi-expert. In A. Bogner, B. Littig, & W. Menz (Eds.), *Research methods series: Interviewing experts* (First ed., pp. 81–97). Hampshire, United Kingdom: Palgrave MacMillan. Retrieved from https://link.springer.com/book/10.1057/9780230244276

Pierre, J., & Peters, B. G. (2020). Different Ways to Think About Governance. In *Governance, politics and the state* (Second ed., pp. 1–18). London: Springer Nature Limited.

Prins, C. (2020). Aansprakelijkheid voor AI-systemen. *Nederlands Juristenblad*, *95*(41), 3141.

Qu, S. Q., & Dumay, J. (2011). The qualitative research interview. *Qualitative Research in Accounting and Management*, *8*(3), 238–264. doi: 10.1108/11766091111162070

Raghavan, M., Barocas, S., Kleinberg, J., & Levy, K. (2019). Mitigating bias in algorithmic hiring: Evaluating claims and practices. , 1–24.

Rahwan, I. (2018). Society-in-the-loop: programming the algorithmic social contract. *Ethics and Information Technology*, *20*(1).

Richardson, R., Chan, A., Kak, A., Diaz, A., Samant, A., Green, B., ... Zapiler, S. (2019). *Confronting Black Boxes: A Shadow Report of the New York City Automated Decision System Task Force* (Tech. Rep. No. December). AI Now Institute. Retrieved from https://ainowinstitute.org/ads-shadowreport-2019.pdf

Richardson, R., Schultz, J. M., & Crawford, K. (2019). Dirty Data, Bad Predictions: How Civil Rights Violations Impact Police Data, Predictive Policing systems, and Justice. *New York University Law Review*, *94*(15), 15–55.

Richardson, R., Schultz, J. M., & Southerland, V. M. (2019). *Litigating Algorithms 2019 US Report: New Challenges to Government Use of Algorithmic Decision Systems* (Tech. Rep. No. September). New York City: AI Now Institute. Retrieved from https://ainowinstitute.org/litigatingalgorithms-2019-us.pdf

Simplilearn. (2020). *Supervised vs Unsupervised vs Reinforcement Learning | Machine Learning Tutorial | Simplilearn*. Retrieved from https://www.youtube.com/watch?v=1FZ0A1QCMWc{&}ab{_}channel=Simplilearn

Snyder, H. (2019). Literature review as a research methodology: An overview

and guidelines. *Journal of Business Research*, *104*, 333–339. doi: 10.1016/j.jbusres.2019.07.039

Sousa, W. G. D., Melo, E. R. P. D., Bermejo, P. H. D. S., Farias, R. A. S., & Gomes, A. O. (2019). How and where is artificial intelligence in the public sector going? A literature review and research agenda. *Government Information Quarterly*, *36*(4), 101392.

Streitz, N., Charitos, D., Kaptein, M., & Böhlen, M. (2019). Grand challenges for ambient intelligence and implications for design contexts and smart societies. *Journal of Ambient Intelligence and Smart Environments*, *11*(1), 87–107. doi: 10.3233/AIS-180507

Torraco, R. J. (2005). Writing Integrative Literature Reviews: Guidelines and Examples. *Human Resource Development Review*, *4*(3), 356–367. doi: 10.1177/1534484305278283

Valtiovarainministeriö - Finansministeriet. (2019). *#AuroraAI - Let your digital twin empower you.* Ministry of Finance - Finland. Retrieved from https://www.youtube.com/watch?v=A2{_}hlJrEiWY{&}ab{_}channel=Valtiovarainministeri{ö}-Finansministeriet

Van Der Voort, H. G., Klievink, A. J., Arnaboldi, M., & Meijer, A. J. (2019). Rationality and politics of algorithms. Will the promise of big data survive the dynamics of public decision making? *Government Information Quarterly*, *36*(1), 27–38.

van der Linde, I. (2021). Het Inlichtingenbureau - Ze weten alles van je. *De Groene Amsterdammer*. Retrieved from https://www.groene.nl/artikel/ze-weten-alles-van-je

van Dorp, J. (2020). *De SyRI zaak en de Amsterdamse situatie Publicaties Autoriteit Persoonsgegevens & Europese Commissie.* Retrieved 2021-05-02, from https://solv.nl/blog/amsterdam-bestrijdt-airbnb-met-algoritme/

van Eck, M., Bovens, M., & Zouridis, S. (2018). Algoritmische Rechtstoepassing in de Democratische Rechtsstaat. *Nederlands Juristenblad*, *93*(40), 3008–3017.

Veale, M., & Brass, I. (2019). Administration by Algorithm? Public Management meets Public Sector Machine Learning. In K. Yeung & M. Lodge (Eds.), *Algorithmic regulation* (First ed., pp. 1–30). Oxford, United Kingdom: Oxford University Press. Retrieved from https://papers.ssrn.com/sol3/papers.cfm?abstract{_}id=3375391

Veale, M., Van Kleek, M., & Binns, R. (2018). Fairness and accountability design needs for algorithmic support in high-stakes public sector decision-making. In *Conference on human factors in computing systems - proceedings* (Vol. 2018-April). Montréal, Canada. doi: 10.1145/3173574.3174014

Verdiesen, I., Santoni de Sio, F., & Dignum, V. (2020). Accountability and Control Over Autonomous Weapon Systems: A Framework for Comprehensive Human Oversight. *Minds and Machines*, *31*(1), 137–163.

Whittaker, M., Crawford, K., Dobbe, R., Fried, G., Kaziunas, E., Mathur, V., ... Schwartz, O. (2018). *AI Now Report 2018* (Tech. Rep. No. December). AI Now Institute. Retrieved from https://ainowinstitute.org/AI{_}Now{_}2018{_}Report.pdf

Wieringa, M. (2020). What to account for when accounting for algorithms: A systematic literature review on algorithmic accountability. In *Proceedings of the 2020 conference on fairness, accountability, and transparency* (pp. 1–18). doi: 10.1145/3351095.3372833

Wirtz, B. W., Weyerer, J. C., & Geyer, C. (2019). Artificial Intelligence and the Public Sector—Applications and Challenges. *International Journal of Public Administration*, *42*(7), 596–615. doi: 10.1080/01900692.2018.1498103

Wirtz, B. W., Weyerer, J. C., & Sturm, B. J. (2020). The Dark Sides of Artificial Intelligence: An Integrated AI Governance Framework for Public Admin-

istration. *International Journal of Public Administration*, *43*(9), 818–829. Retrieved from https://doi.org/10.1080/01900692.2020.1749851 doi: 10.1080/01900692.2020.1749851

Witteman, J. (2020). *De belastingbetaler was onvoorspelbaar, maar nu is er AI.* Retrieved from https://www.volkskrant.nl/nieuws-achtergrond/de-belastingbetaler-was-onvoorspelbaar-maar-nu-is-er-ai{~}ba6ff6c5/

Yeung, K. (2020). *Recommendation of the Council on Artificial Intelligence* (Vol. 59; Tech. Rep. No. 1). OECD. Retrieved from https://legalinstruments.oecd.org/en/instruments/oecd-legal-0449 doi: 10.1017/ilm.2020.5

Yin, R. K. (1994). Case Study Research: Design and Methods. In *Case study research: Design and methods* (Second Edi ed., Vol. 5, pp. 1–53). Thousand Oaks: SAGE Publications. doi: 10.1016/0002-9149(74)90005-8

Yin, R. K. (2018). *Case Study Research and Applications* (Sixth ed.). Thousand Oaks: SAGE Publications. Retrieved from https://us.sagepub.com/en-us/nam/case-study-research-and-applications/book250150

Ziewitz, M. (2016). Governing Algorithms: Myth, Mess, and Methods. *Science, Technology, & Human Values*, *41*(1), 3–16.

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

Appendix A

Preliminary Discussions and Visited Events

Table A.1 contains information about multiple sources consulted to get an idea of what knowledge about the risks of public sector AI still lacks. Most of these were discussions with experts from either TU Delft or the CoA. The experts' names are undisclosed due to privacy reasons. The table also includes the conferences and meetings visited to get an idea of current developments and ideas concerning the topic.

Table A.1: Information about the preliminary talks and visited events

Date	Who or what	Expertise	Organisation
4-2-2021	Anonymous.1	Governance , AI	TU Delft
5-2-2021	Anonymous.2	AI development	City of Amsterdam
5-2-2021	Anonymous.3	Governance general	TU Delft
8-2-2021	Anonymous.4	organisational consulting	City of Amsterdam
8-2-2021	Anonymous.5	AI development	City of Amsterdam
9-2-2021	Anonymous.6	Filosophy of technology	City of Amsterdam, TU Twente
9-2-2021	Conference "Nederland Digitaal"	Digital government	Dutch government
10-2-20211	Anonymous.7	AI for economic department	City of Amsterdam
11-2-2021	Anonymous.8	Governance, AI	TU Delft
11-2-2021	AI Tech Demo	AI development	City of Amsterdam
12-2-2021	Anonymous.9	Governance, AI	City of Amsterdam
15-2-2021	Anonymous.10	Governance, legal	City of Amsterdam
15-2-2021	Anonymous.11	AI development	City of Amsterdam
23-2-2021	Anonymous.12	AI development	City of Amsterdam
1-2-2021	Interview with AI developer	Risk mitigation AI	ALLAI

Appendix B

Generic Interview Information

This appendix presents some generic information about the conducted semi-structured expert interviews in table B.1. The table does not contain names but interview IDs for privacy reasons and to make sure the interviewees felt free to speak. For the same reason their professional position by time of their involvement with the model is left out of the table. The information table does contain the dates and locations of the interviews. Every interview was conducted online through Microsoft Teams due to the Covid-19 pandemic. All interviews are held in Dutch due to the daily language use in municipal practice, except for interview I8. The raw interview data sets are not included with this report, because the transcriptions are too lengthy. These can be obtained with the author on request.

Table B.1: Information about the interviews and interviewees

Interview ID	AI System	Date Interview (2021)	Interview Location
I1	Reporting issues in public space	March 24	Online - MS Teams
I2	Reporting issues in public space	March 31	Online - MS Teams
I3	Reporting issues in public space	March 31	Online - MS Teams
I4	Automated parking control	April 6	Online - MS Teams
I5	Automated parking control	April 13	Online - MS Teams
I6	Automated parking control	April 16	Online - MS Teams
I7	Illegal holiday housing rental	April 7	Online - MS Teams
I8	Illegal holiday housing rental	April 13	Online - MS Teams
I9	Illegal holiday housing rental	April 15	Online - MS Teams

Appendix C

Interview Questions

This appendix contains the interview questions for the semi-structured interviews. They are semi-structured in the sense: the themes for the interviews must be touched upon in every interview, but the exact questions might differ. Questions presented in this appendix are used to fall back on if the interview does not really take off for a specific interview theme. Most interviews are in Dutch and one is in English. The questions are grouped by interview themes. Questions in italics are preferred questions over the others.

C.1 Interview questions

Problem discovery

- Problem identification: What is the societal issue in Amsterdam which led to a demand for the AI model?
- Problem framing and scope: How did you formulate the problem which must be adressed by the AI model?
- Who is involved with the exploration and formulating of the problem?

Stakeholders within organisation involved with AI model

- Which actors were involved with the demand for and development of the AI model?
- What was the division of tasks and did you consider the cooperation to be good?
- Do you think that the different stakeholders within Amsterdam had the same goal for the AI model?
- Do you think that the different stakeholders within Amsterdam all understood the functioning of the AI model and involved risks?
- *Who evaluates the developers team and based on what - in other words, does the protection from citizen risks play a rol in the evaluation?*
- *How does the developers team experience political involvement with their work - is there any steering, do others understand what you do?*
- *Is there enough time, human capacity and financial capacity to come up with sufficient technological solutions?*

Model goal

- Have you considered other options besides this AI model?

- Is the AI model ony used for the goals as stated in advance?
- Based on what factors do you evaluate the success of this AI model?
- What is the AI model capable of doing and what not?

Stakeholders influenced by AI model

- *Are there subgroups of citizens which might experience more negatives of this AI model?*
- Do you think that the model has multiple meanings for different groups of users?
- *Is there a way for citizens to object against the system's decisions or participate in the decisions about the model?*
- *Are there aspects which are important to citizens for this AI model?*

Citizen risks caused by model

- *Which citizen risks do you foresee for this AI model?*
- *How did these risks become clear for you? Were they known in an early stage, or only when the model was implemented?*
- *How do you take the citizen perspective into consideration to determine risks?*
- *What are ways to identify citizen risks for this AI model?*
- *Which risks do you consider to be the most important to prevent or control?*
- Were these risks already known to you in advance?
- *Where do you think these citizen risks originate from?*
- *Which controls do you apply to control or prevent the risks?*
- *Do you think that specific agents have to be responsible for these controls, and if yes, who?*
- Which effects might be caused by faulty classifications by the model?

Considerations of citizen risks and control measures

- *Are risks caused by the model compared with benefits of the model?*
- *If yes; who makes this comparison and if no; do you think it should be the case and who should be responsible?*
- *Are there go/no-go's for the project based on the citizen risks?*
- *Can a risk be too unavoidable or too high impact for the model to keep going and when is a risk acceptable?*
- *Who makes this judgement and based on what?*

Appendix D

Validation Conceptual Model

This appendix presents all methods used for the validation of the conceptual vulnerabilities model. The method is presented in tabular forms in order to be somewhat manageably-sized. First, a validation based on the extensive and varying literature sources used for the model is presented in table D.1.

The next tables concern the validation based on the question: do case-specific vulnerabilities of Amsterdam's AISs found during the case study fit the generic model developed using the literature as just presented? Table D.2, table D.3, and table D.4 therefore present the downsides of the AISs in the case study. These downsides are assigned an ID to use for the next tables.

Table D.5, table D.6, and table D.7 assign the downsides to a model context and specific vulnerability. In this way, the validity of the generic model is checked by comparing it with its potential to interpret real-life cases.

Table D.1: (Semi-)literature consulted for the conceptual model of vulnerabilities

Title	Reference list	Theme
Probabilistic machine learning and artificial intelligence	(Ghahramani, 2015)	Computer Science
The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning	(Corbett-Davies & Goel, 2018)	Computer Science
Conceptual complexity and the bias/variance tradeoff	(Briscoe & Feldman, 2011)	Computer Science
Bias in Computer Systems	(Friedman & Nissenbaum, 1996)	Computer Science
Inverse Reward Design	(Hadfield-Menell et al., 2017)	Computer Science
Hard Choices in Artificial Intelligence.	(?, ?)	Computer Science & Ethics
Ethics as a service: a pragmatic operationalisation of AI Ethics	(Morley et al., 2021)	Computer Science & Ethics
Ethics Guidelines for Trustworthy AI	(AI HLEG, 2019)	Computer Science & Ethics
Accountable Algorithms	(Kroll et al., 2017)	Computer Science & Law
Dirty Data, Bad Predictions: How Civil Rights Violations Impact Police Data, Predictive Policing systems, and Justice	(Richardson, Schultz, & Crawford, 2019)	Computer Science & Law
Thousands of Criminal Cases in New York Relied on Disputed DNA Testing Techniques	(Kirchner, 2017)	Computer Science & Law
Algoritmische rechtstoepassing in de democratische rechtsstaat	(van Eck et al., 2018)	Computer Science & Law
Criminal Law and Technology in a Data-Driven Society	(Hildebrandt, 2015)	Computer Science & Law
Is Intellectual Property Law Ready for Artificial Intelligence?	(Gervais, 2020)	Computer Science & Law
Confronting Black Boxes: A Shadow Report of the New York City Automated Decision System Task Force	(Richardson, Chan, et al., 2019)	Computer Science & Law, Computer Science & Public Administration
Litigating Algorithms 2019 US Report: New Challenges to Government Use of Algorithmic Decision Systems	(Richardson, Schultz, & Southerland, 2019)	Computer Science & Law, Computer Science & Public Administration
Fraudeopsporingssysteem SyRI schendt mensenrechten, overheid moet ermee stoppen	(Huisman, 2020)	Computer Science & Law, Computer Science & Public Administration
What to account for when accounting for algorithms	(Wieringa, 2020)	Computer Science & Law, Computer Science & Public Administration
Fairness and accountability design needs for algorithmic support in high-stakes public sector decision-making	(Veale et al., 2018)	Computer Science & Law, Computer Science & Public Administration
Deze zoektocht naar 'boodschappenfraude' is ondemocratisch	(Dijkstra, 2021)	Computer Science & Law, Computer Science & Public Administration
City in Code: The Politics of Urban Modeling in the Age of Big Data	(Johnson, 2020)	Computer Science & Philosophy
Artificial Intelligence and the Public Sector - Applications and Challenges	(Wirtz et al., 2019)	Computer Science & Public Administration
Aandacht voor algoritmes	(Algemene Rekenkamer, 2021)	Computer Science & Public Administration
Litigating algorithms: challenging government use of algorithmic decision systems	(AI Now Institute, 2018)	Computer Science & Public Administration
Will Algorithms Blind People? The Effect of Explainable AI and Decision-Makers' Experience on AI-supported Decision-Making in Government	(Janssen et al., 2020)	Computer Science & Public Administration
How and where is artificial intelligence in the public sector going? A literature review and research agenda	(Sousa et al., 2019)	Computer Science & Public Administration
Het Inlichtingenbureau - Ze weten alles van je	(van der Linde, 2021)	Computer Science & Public Administration
How to recognize AI snake oil	(Narayanan, 2019)	Computer Science & Society
Concrete Problems in AI Safety, Revisited Anonymous	(Dobbe & Raji, 2019)	Computer Science & Society
Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy	(O'Neill, 2016)	Computer Science & Society
Mitigating Bias in Algorithmic Hiring: Evaluating Claims and Practices	(Raghavan et al., 2019)	Computer Science & Society
A Broader View on Bias in Automated Decision-Making: Reflecting on Epistemology and Dynamics	(Dobbe et al., 2018)	Computer Science & Society
Can You Trust Navigation Apps During a Major Emergency?	(Mak, 2017)	Computer Science & Society
AI in Context: The Labor of Integrating New Technologies	(Mateescu & Elish, 2019)	Computer Science & Society
Algorithmic governance: Developing a research agenda through the power of collective intelligence	(Danaher et al., 2017)	Computer Science & Society
Grand challenges for ambient intelligence and implications for design contexts and smart societies	(Streitz et al., 2019)	Computer Science & Society
Society-in-the-loop: programming the algorithmic social contract	(Rahwan, 2018)	Computer Science & Society, Computer Science & Ethics
Linear Regression Continued, Multiple Linear Regression	(Dobbe, 2020b)	Data Science
Social Data: Biases, Methodological Pitfalls, and Ethical Boundaries	(Olteanu et al., 2019)	Data Science
Bias in Predictie, Aannames voor Predictie	(Dobbe, 2020a)	Data Science
System and Actor Perspectives on Sociotechnical Systems	(de Bruijn & Herder, 2009)	System Science

Table D.2: Downsides of using the AI system for Reporting issues in public space (RI)

Negative aspects of case: Reporting issues in public space (RI)	**I1**	**I2**	**I3**	**Aspect ID**
Bias in favor of outspoken citizens who frequently make reports	x	x		RI.1
Certain need for AI system which was never intended to use for large-scale municipal opertations like these, concerns about controllability	x	x		RI.2
Citizen awareness about this AI system might only be there for a very small part of the population	x			RI.3
Citizens are rewarded for bad behaviour: they dump bulky waste in the streets and file reports to let the municipality take it away	x			RI.4
Citizens sometimes leave their personal information with the reports, which always brings a privacy risk	x		x	RI.5
Developers want to strive for maximum stability of the codes whereas domain users want to increase functionality of the system, trade-off due to limited time	x			RI.6
Human agents blame the AI model for mistakes they wittingly or unwittingly make		x		RI.7
Impossible to correctly handle multiple-issues-in-one reports for the system without human intervention	x			RI.8
Lack of development experts which lead to compromises on AI system quality and strong dependency on individual experts		x		RI.9
Less accuracy of model classifications for public space issues which are less often reported or occur less frequently	x	x		RI.10
Machine Learning potentially contributes to a feedback loop which leads some reports to be handled increasingly less frequent		x		RI.11
New forms of public space inconveniences, like oak processionary caterpillars, repeatedly assigned to wrong category and hard to adapt the AI model		x	x	RI.12
Objective state of the public space does not match distribution of public issue reports; reports should only be interpreted as ask for help	x	x		RI.13
Overconfidence in model performance which leads to classification errors which are not corrected anymore		x		RI.14
Some subgroups of citizens on average do not know how to make reports about public space issues as well as others	x	x		RI.15
Some subgroups on average file more reports which are not recognised or misclassified by the text classifier, for example due to language issues	x	x	x	RI.16
Sudden interference with priorities for system development from alderman, which serves her goals but might not be of interest for complete population	x			RI.17
System provides citizens with unwanted opportunities to make reports about others, for example in light of the Covid-19 measures	x	x	x	RI.18
System should optimise for citizen satisfaction but this so far has not happened		x		RI.19
Trade-off between increasing classifier speed by making the system key-word driven and minimising bias by sticking to historical data as much as possible	x	x		RI.20

Table D.3: Downsides of using the AI system for Automated parking control (PC)

Negative aspects of case: Automated parking control (PC)	**I4**	**I5**	**I6**	**Aspect ID**
AI systems are developed with lacking knowledge of what is necessary for daily operations, outcomes are not useful for street parking department			x	PC.1
Citizens from high-income areas better know how to object against scanning car fines and therefore more often escape penalising		x		PC.2
Citizens game the system by e.g. parking on sidewalks or covering license plates to avoid being scanned without having to pay		x		PC.3
Data leakages have occurred	x			PC.4
Dependency on outsourcing partners for technology development - no oversight from municipal agents	x	x		PC.5
Due to data restrictions, need to use a proxy for livability for model outcome, in this case willingness to pay. Might not be optimal for livability of city for citizens			x	PC.6
Lacking capacity to do surroundings photo checks for all scanned cars without parking permission			x	PC.7
Lot of data involved: can be traced back to individuals if connections with other databases are made	x	x		PC.8
No human decision-maker which recognises the same car being wrongly parked multiple days in a row leads to huge flow of incoming fines for unknowing citizens	x	x		PC.9
Outsourcing partner determines car routing and surroundings checks - no steering from municipality apart from KPIs and checklists	x	x	x	PC.10
Outsourcing partner does not know whether data files have to be kept or deleted	x			PC.11
Privacy infringements, or at least feeling as such, for citizens due to scanning and photographing public space	x			PC.12
Scanning cars provide new opporunities for other forms of citizen control, for example by police	x			PC.13
Unfair treatment of citizen subgroups: cars in some areas are checked more often than others		x		PC.14

Table D.4: Downsides of using the AI system for Illegal holiday rental housing risk (HR)

Negative aspects of case: Illegal holiday rental housing risk (HR)	**I4**	**I5**	**I6**	**Aspect ID**
Algorithmic system makes classifications based on business rules which are not formally determined or described			x	HR.1
Citizens get confronted with unduly fraud checks which might be based on AI model decisions			x	HR.2
Democratically elected officials from city council and aldermen do not have insights into the working of the AI system and its risks			x	HR.3
Draw conclusions about housing fraud based on input information that might not completely be relevant for this outcome	x			HR.4
Hard to check whether the model actually works due to changes in application domain (Covid-19 outbreak)	x	x	x	HR.5
Hard to detect bias because it is hard to define the relevant societal subgroups in real life		x		HR.6
Impact of the model increases if more complaints are made because bottom-of-the-list complaints are not handled anymore, no clarity if this model impact is still acceptable			x	HR.7
Lacking clarity about which municipal department should be responsible for the protection of personal data and IT systems	x			HR.8
Lacking knowledge of AI and algorithms within the municipality		x		HR.9
Model biases as reflections of societal biases		x		HR.10
Model seems to be a means to an end, but in practice makes business rules of the department explicit			x	HR.11
No substantive arguments for all relationships between input factors and model outcomes		x	x	HR.12
Potential model bias or discrimination towards certain societal subgroups	x		x	HR.13
Responsiblities for different aspects of the system divided over multiple municipal departments, which comes at the cost of quality of the system			x	HR.14
Sensitive personal data or proxies for sensitive personal data involved	x		x	HR.15
Team members have access to data sources to which they should not have access	x			HR.16
Team members who are actually data scientists have to become AI developers without extra training	x			HR.17
Using the prioritisation model leads to extensive protocoling of human workers			x	HR.18

Table D.5: Interpreting the mentioned case-specific downsides as generic vulnerabilities for Reporting issues in public space (RI)

Aspect ID	Conceptual model context	Vulnerability
RI.5	AI model	Sensitive and personal information in available data
RI.6	AI model	Acceptability of model implementation
RI.10	AI model	Accuracy of decisions, Training data gaps
RI.16	AI model	System malfunctions for subgroups of citizens
RI.19	AI model	Computational systems are biased towards quantifiable factors
RI.20	AI model	Zero-sum game and trade-offs during modeling
RI.8	Model deployment	Removing humans from decision-making can be a compromise on quality
RI.12	Model deployment	Model not equipped for changing reality
RI.13	Model deployment	Incorrect input-output relationship
RI.14	Model deployment	Lack of model understanding within the organisation
RI.9	Political-administrative	Inadequate financial resources allocated to AI development teams
RI.1	Societal	Contribution to existing social disparities
RI.3	Societal	Contribution to existing social disparities
RI.4	Societal	Reward hacking
RI.15	Societal	Contribution to existing social disparities
RI.18	Societal	Reward hacking
RI.11	Societal	Unexpected or unwanted societal feedback loops
RI.2	-	-
RI.7	-	-
RI.17	-	-

Table D.6: Interpreting the mentioned case-specific downsides as generic vulnerabilities for Automated parking control (PC)

Aspect ID	Conceptual model context	Vulnerability
PC.6	AI model	Computational systems are biased towards quantifiable factors
PC.8	AI model	Sensitive and personal information in available data
PC.11	AI model	Incomplete or incorrect knowledge of Model deployment from modelers
PC.1	Model deployment	Incorrect input-output relationship
PC.9	Model deployment	Removing humans from decision-making can be a compromise on quality
PC.5	Political-administrative	Shifts in discretionary power public agents
PC.10	Political-administrative	Shifts in discretionary power public agents
PC.2	Societal	Contribution to existing social disparities
PC.3	Societal	Reward Hacking
PC.12	Societal	Privacy infringements
PC.13	Societal	Privacy infringements
PC.14	Societal	Contribution to existing social disparities
PC.4	-	-
PC.7	-	-

Table D.7: Interpreting the mentioned case-specific downsides as generic vulnerabilities for Illegal holiday rental housing risk (HR)

Aspect ID	Conceptual model context	Vulnerability
HR.6	AI model	Incomplete or incorrect knowledge of model deployment from modelers
HR.13	AI model	Objective function vulnerabilities
HR.15	AI model	Sensitive and personal information in available data
HR.17	AI model	Flawed knowledge or skills within organisation for AI development
HR.4	Model deployment	Incorrect input-output relationship
HR.5	Model deployment	Model not equipped for changing reality
HR.7	Model deployment	Model not equipped for changing reality
HR.9	Model deployment	Lack of model understanding within the organisation
HR.12	Model deployment	Incorrect input-output relationship
HR.18	Model deployment	Negative impact on workforce
HR.1	Political-administrative	Shifts in discretionary power public agents
HR.8	Political-administrative	Disconnection between AI systems and organisational oversight
HR.11	Political-administrative	Governments adjust their operations to model, rather than other way around
HR.14	Political-administrative	Disconnection between AI systems and organisational oversight
HR.2	Societal	Unverifiable and faulty thresholds
HR.3	Societal	Increasing information asymmetry between public agencies and citizens
HR.10	Societal	Contribution to existing social disparities
HR.16	-	-