Elucidating a 'black-box' transcends explaining the algorithm

Exploring Explainable AI (XAI) as a way to address AI implementation challenges in the Dutch public sector

Abhiramini Rajiv

April 2023

Master of Science Thesis

in Metropolitan Analysis, Design and Engineering

Wageningen University (WUR) and Delft University of Technology (TUD)

Student numbers:	1047832 (WUR) 5435722 (TUD)
Thesis Committee:	Dr. Luciano Cavalcante Siebert (TUD, Supervisor) Prof. Dr. Alessandro Bozzon (TUD, Supervisor) Dr. Tim van Emmerik (WUR)
Additional Supervisors:	Mireia Yurrita Semperena (TUD) Sietze Kuilman (TUD) Steven Vethman (TNO) Ilina Georgieva (TNO)

Abstract

Responding to the trend of increasing use of artificial intelligence (AI), we need to ensure applications of AI are designed, implemented, utilised and evaluated in a careful manner. Explainable AI, or XAI, meaning; - given a certain audience, the details and reasons of both technical processes of the algorithm-support system and the reasoning behind the system to make its functioning clear or easy to understand - is one of the ways to responsibly design and implement AI systems. This research looks into AI-supported public decision-making processes in the Netherlands and the role and possible contribution of XAI in such a context. To this end, I conducted a mixed-method qualitative study; interviewing sixteen respondents from three key-actor groups within two Dutch national public sector executive bodies, additionally performing three observations and document-analysis. Differentiating between different phases of an AI system's implementation life-cycle, the study unveils how the respective actors - *managers*, *data scientists* and *domain experts/(potential) AI users* - encounter various challenges in bringing an AI system from idea to production. The empirical findings show that many AI systems, whilst technically developed, are not deployed or adopted by the wider organisation. The study discerns the challenges hindering the AI implementation process from an organisational, human and technical point of view. Moreover, the study highlights the need to approach XAI from a multi-purpose, multi-actor perspective; both acknowledging that various actors need different kinds of explanations, but also bridging different respective professional worldviews to apprehend one another. XAI is often seen as a one-size-fits-all solution for various implementation challenges, however the study shows that certain challenges need to be addressed at least beyond traditional ways of XAI from a computer science perspective, and perhaps beyond XAI all together. As such, the insights of this thesis contribute to generating a more realistic idea about the opportunities and limitations of XAI, within real-world AI implementation processes in the public sector.

Key-words: Artificial Intelligence, Explainable AI, Public Sector, Algorithmic Decision-making, Societal Impact.

Contents

1 Introduction

In an era where society is larded with the ubiquitous promise of Artificial Intelligence (AI), the public sector seeks to keep abreast. AI-based solutions provide an opportunity to incorporate innovation, quantified methods and data into endeavours addressing deeply rooted problems. That said, if not considered carefully, the implementation of AI systems may have far-reaching and possibly adverse consequences. This is true for the usage of any AI system really, but it becomes even more crucial dealing with societally sensitive topics in the public domain. Another layer of complexity is added due to the nature of the public good and its diverse stakeholders. Using algorithms for public sector decision-making implies they need to aid 'the common interest'. Yet such a common interest is by no means unambiguous nor even 'common', in a field consisting of inherently divergent views and often dealing with sensitive issues.

The most recent instillment of awareness in the Netherlands was raised during the childcare benefits scandal, also known as '*de toeslagenaffaire*'. An evaluation report by Amnesty International, titled "Xenophobic machines – Discrimination through unregulated use of algorithms in the Dutch childcare benefits scandal", exemplifies the possible dangers of irresponsible use of algorithms [1], and drew attention to the topic. Following these and other calamities due to the uncareful use of models, such as the painfully biased set of Google photos of African-American people labelled as 'Gorillas', or the false positive rates for people of colour being disproportionately high [2, 3], the necessity and importance of a responsible use of AI has become clear, underscored once again by the European GDPR (General Data Protection Regulation) in 2018 for collection, storage and use of personal information and the European Commission currently proposing a European law on AI, the AI Act, as a first law on AI by any major regulatory body [4, 5].

1.1. Problem Statement

Responding to the trend of increased technological advancement including in the public domain, one ought to account for possible unfair and/or unintended outcomes, even when society's best intentions are at heart [6]. By now, most practitioners and researchers agree that AI systems used in the public domain need to adhere to a minimum set of values[7]. Much heard values are 'Responsibility', 'Explainability', 'Contestability', 'Accountability' and 'Trustworthiness' [8]. Although this is generally acknowledged, the difficulty arises when one tries to translate these, even conceptually debatable but at least righteous, values to technical and organisational workflows [9].

Taking two national public sector executive bodies in the Netherlands as case studies, this thesis aims to disentangle public sector AI implementation processes. Conversations with municipalities and national governmental organisations leading up to this research, showed that public sector bodies are struggling with the design and implementation of such measures, manifesting challenges, uncertainties and trade-offs during the incorporation of AI-models within the organisation. Literature confirms that multidisciplinary research on the use of AI for public governance, including effective implementation and the management of risks is necessary [10, 11]. In quest of reaching societally beneficial decisions aided by algorithms then, the various values mentioned above could or ought to be considered. The scope of this research focuses on 'Explainable AI', 'Explainability' or 'XAI', as one of those values. One could pinpoint and study differences between the three concepts, but this is beyond the scope of the research. For the remainder of this thesis, Explainable AI and XAI will be used interchangeably, and explainability is referred to within the context of AI. It is reasoned that one cannot responsibly use algorithms for public sector decision support if there is no- to insufficient understanding of the AI system. Explanations offer a means to approximate such understanding [12]. Therefore, arguably, incorporating some form of Explainable AI becomes a prerequisite to firstly improve understanding of the systems to one's best ability and, secondly, extrapolate how these could be deployed beneficially though at the same time minimising possible harm.

Having said the above, the concept of Explainable AI is not a clear-cut one. Literature has various definitions on the topic, differing per discipline as well. Most often the concept is discussed from a computer scientist's point of view, describing that "Explainability provides explanations on the functioning of a model" [13]. However besides a 'technical' view on Explainable AI, one can take a more 'human-centric' or 'system-centric' approach, reasoning that explaining the functioning of a model alone is not enough to fully understand the origin, use nor consequences of a model. From a socio-technical, system's perspective, Explainable AI raises broader questions as well, such as what the explanation means for different audiences, why certain choices have been made, to what extent the model could be contested or re-designed if deemed fit and what could be the adverse effects of the model's outcomes in practice [14].

1.2. Research Aim

Drawing from the problem statement, the aim of this research is three-fold. Firstly, it is the aim to understand if and how Explainable AI could be of value for Dutch public sector executive bodies. Taking public sector decision-making practices as a focal point, more specifically researching Dutch national executive bodies with a supervisory purpose, it is important to map the current role of both AI as well as Explainable AI within an organisation; understanding the recurring challenges and how these challenges can or cannot be overcome by exploring the possibilities of Explainable AI. Secondly, the aim is to identify and compare the perceptions and needs of different stakeholders during the process. Intuitively it might seem logical that a decision-maker needs different explanations than a data scientist, yet what they would need exactly, and which type of explanations would adhere thereto is less obvious. Within the research context, at least four types of stakeholders are relevant; the *managers* deciding whether and how to include AI on a strategic level, the *data scientists* developing the algorithms, the *decision makers* or *domain experts* using the algorithms and the *decision subjects* affected by the decisions made. Mainly due to the practical difficulty in identifying, finding and reaching the decision subjects, the scope has been limited to the managers, data scientists and domain experts.

Lastly, closely related to the prior two, is the aim to translate the theoretical definition and goals of Explainable AI to their role but also potential contribution in practice. Theoretically, there are many benefits to Explainable AI which ought to be explored. Nevertheless, practice works in unexpected ways, making it meaningful to research how the theoretical benefits and pitfalls abide within the research context. Operationalising Explainable AI from a technical perspective is increasingly done, for instance by using XAI methods such as LIME [15], bLIMEy [16], SHAP [17] or DiCE [18]. A more elaborate and profound disquisition of the concept will be explored in the literature review, but it is most important to note that this research ventures to broaden the concept of Explainable AI to incorporate explanations beyond the model. Therefore, the third research aim is to contribute to the operationalisation of the theoretical purpose and opportunities of XAI from a socio-technical perspective.

In summary, the goal of this research is to 1). understand the value of Explainable AI, 2). as defined by identified key-stakeholders, 3). by making it practical for decision-support systems in Dutch public sector executive bodies. Thereby hopefully contributing to the use of AI systems in a beneficial way, diversifying and bridging between stakeholders where necessary, and translating theoretical ideas into practical suggestions for improvement.

1.3. Research Question

The central research question to reach the different research aims, is formulated as follows:

"From a Manager's, Data scientist's and Domain expert's point of views; What is the role and possible contribution of Explainable AI for responsible implementation of algorithm-aided public decision making practices?"

To answer the main research question, it is relevant to first extricate current practice and its challenges, to know where Explainable AI is already used and how so, but also to know which parts of the process ought to be addressed, whether or not through the use of Explainable AI. Next, we need to know what is understood by 'Explainable AI' and what are the potential benefits. Since Explainable AI is closely related to understanding, it is useful to dig into the different ways of reaching such understanding; both for different stakeholders as well as for the different purposes Explainable AI can serve. Therefore, it is worth elaborating on the required and preferred information needs of involved actors vis-à-vis the AI system they are working with. Finally, reaping the practical findings, the research benefits from studying and exploring strategies for improvement.

The sub steps to answer the main question, are transliterated to four sub-research questions:

1. What are practices and challenges of implementing AI in the public sector?
2. What is the current role of XAI in the public sector?
3. What are the information needs, to improve understanding of the AI system, as defined per stakeholder?
4. What could be the contribution XAI in addressing AI implementation challenges?

1.4. Scientific and Societal Relevance

AI has fervent advocates, ardent adversaries, oblivious users, fearing subjects, indifferent bystanders and probably everything in between. All these stances have some truth to them, not just by the fact that novel technologies always play out differently than expected, but also because the types of AI and their implementation can be so widely differing. What is more, people with their worldviews and different level of knowledge on the topic subsequently showcase different approaches to understanding. The scientific relevance of this research then is to acknowledge these different approaches, view them as valid and aim to broaden a predominantly technical field towards a socio-technical and multidisciplinary one, including but not limited to bridging the gap between the domains of governance and computer science. More so, the scientific value of this work lies in the access and focus on practical cases. The topic of Explainable AI (XAI) has become increasingly popular over the last few years, but mostly driven from an urge to discern technically challenging, non-transparent algorithms [19]. However, models within many public bodies to date are often not as technically complex, even though their implementation and governance are. It is scientifically relevant then to see how the role of XAI could transcend the workings of an algorithm, making it more potentially relevant in wicked practical settings. In other words, it is relevant to go from "Explainable AI" as a algorithm-centric concept to a practice-based operationalised "Explainable AI".

Societally, the relevance of explainable AI implementation is its promulgation in combination with insufficient knowledge on containment of future mishaps. We have already seen that using AI implies the promise of improving decision-making, yet also that it introduces new biases and risks especially when non-quantifiable values are translated to quantifiable inputs for an algorithm. Being on the verge of more widespread AI implementation within the Public Sector, we ought to avert the hurting of people as a result of uncareful implementation of AI. Of course, with the introduction of new technology and models, mistakes will inevitably be made. It is the way these are handled however, the way one learns and makes sure mistakes do not turn into disaster which are at stake here. The contribution of this research then, is the enhancement of understanding of AI systems throughout their initiation, design, implementation and evaluation, with the long-term purpose of understanding how to defy discrimination and potential harms.

1.5. Reading Guide

The chapter hereafter reviews literature on AI within the public sector and Explainable AI, to come to a first understanding of all four formulated sub questions (chapter 2). Next, the methodology of the research will be expanded upon (chapter 3). Interviews and observations are conducted within the two case studies, to get an in-depth understanding of varying perceptions, challenges and opportunities of the different stakeholder groups. The empirical findings are divided into three parts; the AI implementation practices and challenges (chapter 4), the role of XAI and the information needs to make AI more explainable (chapter 5), and the connection of the findings to analyse how XAI could contribute to AI implementation challenges (chapter 6). The discussion reflects upon the research content, context and process, featuring both limitations as well as suggestions for further research (chapter 7). The thesis concludes by formulating an answer to the main research question and bringing together the different insights (chapter 8).

2
Literature Review

This literature review gives an overview of earlier conducted research; to ground the theoretical side of the sub-research questions (section 1.3). First, the usage of AI within public sector practices is discussed, to get a better understanding of the practices we are talking about. Next, the potential hazards accompanying public sector AI implementation are discussed. This is relevant for two reasons; the hazards might pose challenges to implementation practices, thereby being a reason for the increasing call for 'responsibly' implementing AI. Then, following the call for responsible AI, the emergence of Explainable AI as one of the responsible AI values [20], and a disquisition of the concept are deliberated upon. Finally, various types of explanations which could lead to meaningful Explainable AI, are considered. The chapter concludes by identifying trade-offs and controversies around the topic.

2.1. The Use of AI in Public Sector Decision-making

The scale of using Artificial Intelligence (AI) within decision-making, both in public and private sectors, has been expanding rapidly [21]. These practices, also referred to as part of "algorithmic decision-making", look at instances or processes where algorithms make decisions previously made by humans. A few recurring themes can be discerned from discussions around the topic in literature, three of which will be highlighted in the context of this research. Firstly; the nature of the decisions made, subsequently influencing its algorithmic companionship. This includes a subsection on the reasons to use algorithms within public decision-making at all. Secondly, the nature of the algorithms used. And thirdly, how the algorithms are implemented according to literature.

2.1.1. The nature of *public* decision-making

Governments are becoming increasingly data-driven and are looking for ways to use data and algorithms in their decision-making processes, and yet, research into AI within government practices is still scarce [11, 22]. Even though algorithmic decision-making happens within multiple domains and is researched as such, the nature of decision-making in public administration and government is particular and a research domain in itself. There is a long tradition of literature focusing on such decision-making practices[23, 24], emphasising that "they often need to be solved based on incomplete, contradictory, and changing information" [11, p. 480]. The nature of these problems faced by public bodies are often referred to as 'wicked problems'. Decisions to solve such problems then, are not straightforward nor is the process that leads to these decisions.

Simon [25], proposed a theory of "bounded rationality", arguing that the rationality within this domain is bounded by "the intractability of natural decision-making problems, the cognitive limitations of the mind of human decision makers, and the limited time available to make the decision". Lindblom [26] went even further in his claims that no theories are used at all, and that decisions are made quickly based on very little information. To date, many scholars have aimed to narrow down political decision-making processes to rational workflows and schemes, for instance drawing up variations of Easton's System's model [27], Rational Actor Models [28] or Kingdon's Stream/Barrier Model [29]. The adaptions of these models often balance a fine line between making linear schemes less staccato therefore more reflective of the complex reality, at the same time keeping them illustrative to present reality more comprehensibly.

Given the complicated and blurry process of arriving at decisions in the public sector, adding algorithms to the mix does not make the process less complex. One could try to disentangle such novel dynamics within the current public decision-making literature, but these theories only account for 1). political processes and 2). human interactions, thus leaving out automated actors and their implications. What is more, with the rise of internet, citizens increasingly expect services to be provided online and decisions to be provided within a shorter time frame [11]. While the use of the algorithms is increasing as a result of this, the influence of their use for public decision-making is less well understood [10]. Specifically studying the role of algorithms within the context of public decision-making then, seems valuable, as it gives rise to new challenges, trade-offs and might raise a request for amended checks and balances. As a result, the use of AI in the public sector introduces the relevancy for newly coined terms like ‛data- driven government' or ‛algorithmic governance' [30].

We now know that algorithms are increasingly used, including in the public sector. We also know why it is important and useful to look into algorithmic governance additionally to ’regular’ public sector decision-making literature. Before moving towards understanding how such algorithms find their place, the reasoning behind the use of the algorithms at all, seems relevant. Knowing *why* algorithms are introduced, will help contextualise *how* they are implemented later on.

Rationale: Why add algorithms to public decision-making?

AI within the public sector can be used for multiple reasons; to better deal with vast amounts of data, to (partially) automate decision-making processes with the aim of making these processes more efficient or to optimise and innovate certain public functions that could not, or could not sufficiently, be addressed before. The use of algorithms in decision-making is generally embraced for situations in which a large amount of decisions needs to be made on a large scale, within a short time frame [11].

Some say AI can be used to achieve more rational, more accurate or less-discriminatory decisions, though such claims are controversial [22]. Reports about the benefits of AI argue that AI enables organisational employees to reach better decisions, to boost our analytic and decision-making abilities and heighten creativity [31]. A key premise to such claims is that better information will lead to better decision-making [11, 32]. This follows a line of a public administration ideal where decision-making is bureaucratic, highly organised and thorough, closely related to the wish to capture wicked problem decision-making into logical workflows. Terms like ”evidence-based” and ”data-driven” decisions are used to refer to this sub-field of AI-supported decision-making [33]. However, even though such ’rational logic’ in combination with the growing capability of AI technology to undertake more complex tasks are increasing the use of AI in the public sector domain, the notion that reality is more cumbersome is starting to gain ground as well. Indiscriminately adopting AI for public decisions has proven to be susceptible to bias and unable to deal with variety [34, 35], therefore making the question *why* algorithms are being used in a specific context extremely relevant to ask, every single time.

2.1.2. The nature of *algorithms* in public decision-making

How then, do these data-driven and evidence-based decisions materialise? How do algorithms find their way in public sector decision-making? Those questions are addressed in this subsection. Different aspects of the nature of algorithms in public decision-making could be highlighted; this section focuses on two of these aspects. Firstly, how the algorithms (i.e. ’the machines’) are used in interaction with the people using the algorithms. Secondly, the types of algorithms currently predominating the sector.

Automation versus algorithm-support

One of the debates surrounding the nature of algorithms in the public sector, concerns the extent to which algorithms act autonomously versus if they are rather a supporting mechanism for human-made decisions. A general consensus, given the often sensitive and societally relevant topics, is some variation of the latter. That said, less agreement exists on the degree and manner of human- and/or machine involvement.

Roughly speaking, one could distinguish a dichotomy where decisions are based on the outcomes of the model (automation), or based on the human who is assisted in by a model (augmentation or AI-support) [21]. As the progress of AI technology enables to undertake more complex tasks that require cognitive capabilities, such as making tacit judgements, sensing emotion and driving processes, decisions could in theory be increasingly (over)taken by AI systems [21, 36]. Therefore, "with the resurgence of AI, a new human-machine symbiosis is on the horizon and a question remains: How can humans and new artificial intelligences be complementary in organisational decision-making?" [37, p.579]. Jarrahi [37] calls for a new human-machine symbiosis, in which a shifting division of work between machines and humans is described. Originally, such a division of work would be that machines take care best of mundane tasks, leaving humans to do more creative work. However, given increased technological possibilities, the 'task-division' or collaboration is not as simple, thus more advanced ways to describe the mechanism of using AI in tandem with human considerations are researched, such as through Human-Computer Interaction (HCI), Team Design Patterns or Learning Design Patterns (LDPs) [38]. One way to describe the interaction is by talking about comparative advantages by humans and machines respectively; AI being better at overcoming complexities through superior analytical capacity, yet humans being better at dealing with uncertainty and equivocality (negotiate, consensus and build support) in decision-making, through their intuition[37]. As a result, machines ought to depend on humans for subconscious heuristics and both human evaluation as well as facilitation of decision outcomes remains necessary. Such claims are more prevalent if the impact of the decision might be large, i.e. if 'the stakes are high'. Within the public domain, the use of algorithms can differ from more mundane tasks to more sensitive ones. Nevertheless, generally even if the task is considered easy or mundane, unwanted biases, privacy issues, inequality and implementation mistakes are lurking, therefore favouring semi-autonomous decision-making over autonomous ones. In other words "AI systems should be designed with the intention of augmenting, not replacing, human contributions" [37, 584].

Based on the above, even though the exact considerations for using AI and its implementations are context specific, this research is based on the general assumption that an algorithm, 'machine' or computer does not reach a decision on its own in the public sector. Rather, it is to support the decision-making systems already in place, based on the interaction between the model's inputs and outcomes, human interpretations and other information that ultimately lead to a decision. The British Information Commissions' Office (ICO) represents these various components in figure 2.1 [39].

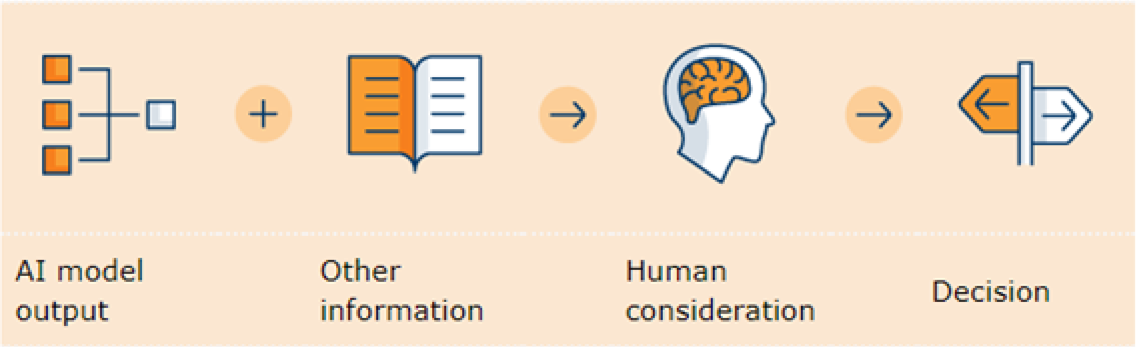

Figure 2.1: ICO's Model - Information - Human interaction to reach decisions [39]

The types of algorithms used

Knowing that the algorithms are generally used to inform and support human-made decisions, it is interesting to know which types of algorithms are contemplated for such a purpose. Here, it is useful to mention that AI is a broad term and its applications in public decision-making then, are varied as well. Janssen et. al [11] distinguish two types of algorithms mainly used in the Dutch public sector decision-support, Machine Learning (ML) and Business Rules (BR). They argue that, where Dutch public sector practices used to predominantly make use of the logic of BR, they are moving towards ML.

However, at the same time they argue that moving from rule based to computational algorithms ought not to happen until the explainability of such algorithms is improved.

The ML algorithms, in short, derive patterns from data, without explicit instructions. For BR, humans define explicit rules based on logic and assumed causal relations. The respective algorithm types, fall under a more generally made distinction between two types of algorithm families [40]. The first contains 'Learning' techniques (which are most predominantly statistical in nature). These are more classically called 'sub-symbolic'. The second contains 'Knowledge Representation' techniques (which are predominantly discrete in nature). These could also be called 'symbolic'. Note that the distinctions are not uncontroversial. Moreover, there exists a growing literature body advocating for the combination of the methods [40]. Exploring this in itself could be an elaborate study, yet, it is beyond the scope of this thesis. Within the scope of the thesis however, is their general application and the effects of different algorithms on explainability.

Knowledge Representation systems, more specifically rule-based systems based on declarative logic, gained popularity in the public sector in the 90s already [21, 41, 42]. These systems are considered to be less advanced or complicated than ML systems, as the rules reduce the need for complex programming and make it relatively easy to understand [11]. Drafted by a human, they are based on current or historic practices, procedures or legislation. They could be seen as a 'representation' of human decision-making. Ideally, the domain knowledge and legislation would be one-to-one translated down to the model, yet in reality going from a given situation to an algorithmic system, there is the step of interpretation which makes it less straightforward than the 'simplicity' of the models might suggest [43]. There are a few more mentioned downsides to rule-based systems, such as the dependence on availability of information. These are important to understand if one were to make a trade-off between different types of models, however exploring them in-depth goes beyond the scope of this thesis.

Since ML implementation processes are less established, more recently introduced and currently topical within the field of study and practice, it seems most interesting and relevant to look into these algorithms. Therefore, even though it is acknowledged that AI can be seen in a broader light, the remainder of this thesis focuses on (Machine) Learning algorithms and their implementation process. Before starting the contents of the following subsection, two considerations are worth mentioning. Firstly, even though research reports that ML and BR are mostly used within the Dutch public domain context, literature thereof is not very extensive. Secondly, the advancement and complexity of ML methods used, differ from organisation to organisation.

Machine Learning: debates and limitations

Algorithms learning by identifying patterns in data sets are increasingly considered in the public sector. Especially ML that uses data to overtake structured human tasks, receives a lot of attention [35, 44]. The focus then often lies on the inputs and outputs of the model, however, a main criticism there is the opacity of the ML [2]. Multiple authors refer to a "black-box' when it comes to ML algorithms, since the internal logic of the algorithm is defined by the model and not by human beings [2, 3, 6, 45, 46]. This form of opacity is explained in the following quote: "While datasets may be extremely large but possible to comprehend and code may be written with clarity, the interplay between the two in the mechanism of the algorithm is what yields the complexity(and thus opacity)" [2, p. 5]. Moreover, fairness, (non)-discrimination and the quality of data are topics which come up discussing the societal impact of such opaqueness. The opacity makes it hard to scrutinise decisions and to ensure that decisions are always made by applying the same logic[11]. The data on top of that is almost mostly biased, due to various institutional practices amongst which the historical paths of the data gathered, the background and culture of the data subjects whose data is gathered, the frames of reference of the decision makers and those gathering data. Inadvertent biases are thus introduced, often reinforcing historical discrimination or reflecting outdated practices where societal or political preferences have changed [47, 48].

ML systems have their controversies and limitations thus, posing plausible effects on their (potential for) explainability. Below, the limitations are summarised, mostly following van Harmelen [40]. The limitations of learning systems are:

- *Opaque*: meaning that is typically difficult to extract a humanly-comprehensible chain of reasons from the output of the system;
- *Data hungry & inheriting bias*: learning systems, especially as they grow more complex, need large training sets. Note that the quality of data matters a lot as well, since bias is inherited from historical data;
- *Limited transfer*: a trained network performing well on one tasks, may perform poor on a new task;
- *Brittle*: even with similar inputs for the same task, the outputs may differ significantly;
- *Limited use of prior knowledge*: the performance of learning systems is based on the data during their training phase, and is not informed by general principles such as general domain knowledge. Therefore, it is difficult to deal with exceptions or unusual cases.

2.1.3. *Implementation and adoption* of algorithms in public decision-making

A few researched examples of implemented public sector decision-making worldwide are Public Employment Services and Medical practices [49–52]. Looking at the Public Employment Services for example, the deployment of algorithms has sparked quite some controversy, to the point that some systems have been removed or the role of AI there within was reduced [50]. Notwithstanding, the implementation of similar systems continues, making it useful to learn from the well- and less-well received practices, before and whilst designing and implementing current AI-supported public decision-making systems.

In the Netherlands specifically, the AI-applications supporting public decision-making range from mundane tasks such as granting permits and calculations of tax returns to more nuanced tasks such as granting social benefits, healthcare practices, admittance of immigrants or customs surveillance[11]. However, Janssen et al. [11] argue that empirical research into the influence of the more nuanced tasks, especially involving complex models, is limited. While there is an increased tendency to use the algorithms, their benefits and challenges are less well understood [11]. The empirical part of this research will dive into those benefits and challenges. To be able to do so structurally, this section underpins the process of establishing AI by understanding their implementation and adoption.

AI implementation, as defined by Damanpour [53, p.217], refers to "the events and actions that pertain to (...) preparing the organisation for its use, trial use, acceptance by the users [and finally] use of the innovation until it becomes a routine feature of the organisation". Basically, AI implementation encompasses all the phases to be gone through, to enable putting it into use by an organisation.

AI implementation life-cycle

This subsection dives into the way AI systems find their way to deployment. Within the AI implementation process, several phases can be identified; one needs to arrange certain things before starting to design an algorithm, one needs to design and develop the algorithmic-system, the system needs to be tested, it needs to be put into use and evaluated. Note that the phases do not linearly follow one another. In an ideal world, one would arrange everything neatly before starting the design of a process, and the tests would be perfect launching right into deployment. In a more practical world, the phases pass by in iterations, and sometimes even intertwine all together. For example, the development or testing of a system might lead to identification of flaws, requiring to go back to the drawing board. Therefore, this study chooses to depict the implementation process as a life-cycle. Note that a life-cycle representation still sells short of capturing the entire complexity at hand, but for analysis' sake and because it does reflect a cyclical and iterative process, it is considered meaningful nonetheless.

Inspired by ICO [54] and Alfrink et al. [55], the following phases are determined (graphical depiction in figure 2.2 below):

- Initiation & pre-conditions;
- Design & development;
- Pilot or Testing;
- Deployment;
- Monitoring & evaluation.

ICO [54] and Alfrink et al. [55] define the first phase as business and use-case development, where the problem and/or improvements are defined and a AI system is proposed. Ex-ante safeguards are put in place to protect against potential harms. The set-up of the use-case and preconditions for a fruitful organisational and modeling environment, are captured in the *initiation & pre-conditions* phase of the AI implementation life-cycle. Next, the distinguished phases of design, data procurement and development are taken together as a second phase *design and development phase* for the purpose of this research. The reason to mention the three elements in one phase, is that in the public sector, they tend to be closely intertwined. In the design phase, the AI business case is turned into design requirements. In the data procurement, input data sets are obtained to train and test the model. In the development, the design requirements are translated to an actual model, fed by the gathered data obtained. Successively, the testing phase, often done in the form of a 'pilot', is worth considering more in-depth in the context of public sector implementation. Pilots and experimentation are considered critical for many novelties in the public sector, including AI applications, to identify and mitigate risks of failure which may prove disastrous in eroding citizen trust [56, 57]. As mentioned earlier, the majority of ML projects in governments are currently pilot applications, rather than deployed ones [58]. The proliferation of innovation labs is a testament to a realised need for experimentation with new technology applications [59, 60]. The idea is that smaller successes enable organisations to mature and build capabilities before undertaking a large-scale AI-driven system-change. Following the testing phase, comes the *deployment phase*; bringing the AI system into production. However, since many of the AI applications have not reached deployment yet, this phase will receive less attention in this thesis. Ultimately in the AI life-cycle, continuous monitoring, and evaluation is considered. Generally in implementation life-cycles, the *monitoring and evaluation phase* is mentioned following the deployment phase. However, I would like to argue that evaluation and learning needs to happen continuously and in a structured manner (optionally, though depending on the context, third party oversight strengthens the evaluation process [55]). Therefore, in the depiction below (figure 2.2), monitoring and evaluation can be read before, throughout and after each of the phases.

AI Adoption

AI adoption adds a layer on top of implementation, ensuring that the AI system is not only deployed but also embedded within the organisation. AI adoption entails "an integration of the new knowledge through creating new capabilities, technologies and training programmes" [61, p.1008]. Besides providing technical solutions then, integration needs to be addressed through socio-technical mechanisms as well. Literature on technology adoption vouches for adoption to be an inherently social and personal developmental process. Straub [62] argues that individuals construct unique yet malleable perceptions of technology, in turn influencing their willingness and ability to adopt the technologies at hand.

Closely related to the aforementioned facets of adoption, Madan et al. [57] distinguish three characteristics that allow for adoption: organisational characteristics making up the environment and surroundings of the AI-adoption context, human characteristics that predispose a person or persons to seek out or shun change and the technical innovation characteristics specific to the use and compatibility of the particular innovation, in this case the algorithmic decision-support model. Looking at Dutch public sector decision-making processes then, and especially analysing the difficulties in reaching embedded and widely-accepted deployment, the AI implementation life-cycle incorporates these three important factors related to adoption in every phase; organisational-, human- and technical factors.[1]

In summary, algorithms are increasingly being used for decision-making. The public context makes decision-making processes urgent and interesting, yet complex in dealing with wicked problems. The systems tend to be algorithm-supported rather than autonomously acting machines, yet the exact form of interaction between the human and the machines differs. The complexity of the models, in this thesis scoped to ML, might have implications for their explainability later on. Even still, from an ethical and societal perspective, it is important to review every implementation process including their adoption, whether the algorithms involved are complicated or not. Literature identifies limited empirical knowledge of the impact of AI on the public sector [10, 11]. Next chapter looks into potential challenges of AI models, before diving into Explainable AI to account for some of these challenges (section 2.3).

[1] Note that one could incorporate other factors (e.g. economic or factors), or specify the three factors distinguished (e.g. subdividing organisational factors into political, legal and policy-related) [52]. However, this is outside the scope of this thesis.

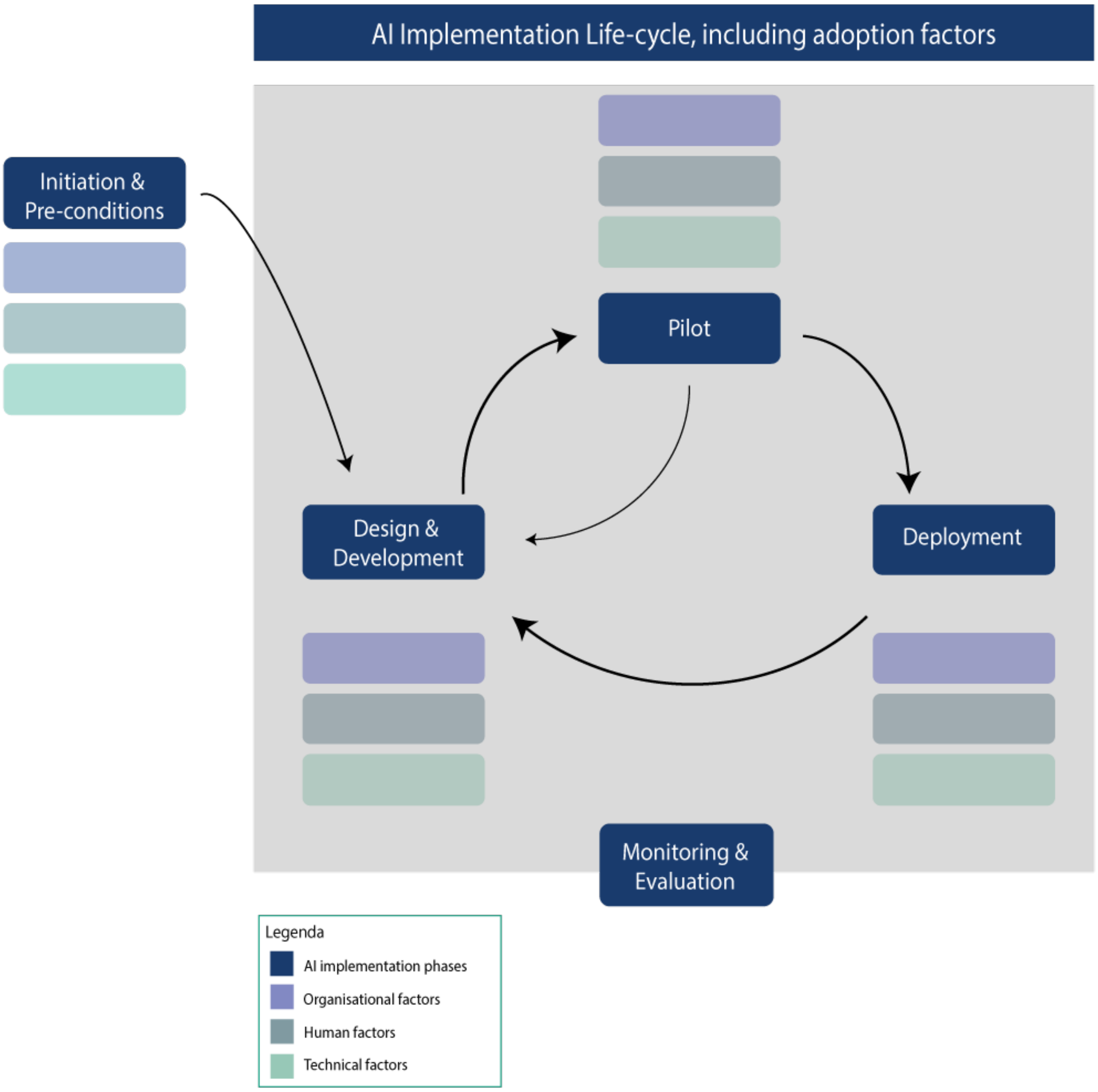

Figure 2.2: AI Implementation life-cycle, as inspired by the phases of ICO[54] & Alfrink et al. [6]. Dark blue represents the life-cycle phases (evaluation being present throughout the life-cycle). The light shades represent the adoption factors to be considered per phase.

2.2. Why Models Might Go Wrong: AI Implementation Challenges

"Nothing is inevitable, until it happens" [63, p.16]

People are, and arguably ought to be, reluctant to adopt systems which are not directly understandable or trustworthy [14, 64]. As AI grows more sophisticated and ubiquitous, the voices warning against its current and future pitfalls grow louder. The recent 'toeslagenaffaire' mentioned in the introduction is just one of the examples of unwanted consequences of the use of models. In a February 2018 paper titled "The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation" 26 researchers from 14 institutions enumerated various sources of harm caused by AI-supported systems, amongst which the risk of increasing socioeconomic inequality and the possibility to falsify data, more commonly known as 'fake news', altering the public debate and the opinions of millions [65].

This section outlines some frequently mentioned hazards, potentially leading to harms of using AI within the public sector specifically. However, not all AI systems lead from hazard to harm of course.

Much of the debate is still clouded by ill-information, ill-understanding and fear but also by boasting, overconfidence or recklessness. This chapter does not have the aim of rejecting or critiquing the use of AI in advance. Rather, it is the aim to map main potential hazards and which (negative) impact they might have, to then know what to pay attention to in further implementation. For the purpose of improved oversight it is useful to categorise the sort of hazards. In accordance with the aim to approach explainable AI beyond the scope of the model, the hazards will be viewed beyond the model as well. Whether in the domain of urban planning, in digitisation or in the field of AI, multiple authors have argued that our systems tend to overemphasise the technological at the expense of looking at social aspects, suggesting socio-technical misalignment [66–68]. This in itself could considered to be a hazard. The following section outlines often mentioned hazards, both from a technical as well as from a socio-technical point of view. This study refers to *socio-technical* when social and technical aspects are intimately related [69]. Note however that the list is not claimed to be exclusive nor exhaustive. Thereby, it is not a clear-cut distinction. Whilst some might view a certain topic as an exclusively technical topic, others might say its a socio-technical topic and vice versa.

2.2.1. Technical-hazards

Technical hazards link back directly to the design, input, internal workings and outputs of the model. Below, two types of technical hazards will be described, closely related to the modeling process, namely: the input data, and inadequate outcomes of the model in the form of false positives or false negatives.

Quality and quantity of data

"AI agents are only as good as the data human put into them"[70, p. 81]. Machine learning and data-driven decision-making in general depend heavily on the quality of the data. Data can be historically biased, due to institutional practices, due to the background and culture of the data subjects whose data are gathered in the past. For instance, software used to predict future criminals was biased against a certain race [70]. The bias comes from the training data that contains human biases. As such, the input of the model, i.e. the model's input data, can lead to historical discrimination reinforcement [48], or a continuation of outdated practices whilst societal values and political preferences might have changed.

Furthermore, the quality of data does not merely lead to potentially biased decisions, it can also lead to inaccurate or inadequate ones. If we look at algorithms predicting risks for instance, the algorithm does not interpret or judge if the output is based on an unwanted or unjust factor, such as the under-representation or over-representation in the data of a certain group, versus if it 'justly' identifies an instance because the risk is really higher. Data fed into the algorithms for learning determine its outcome thus, and ultimately the quality of the data affect the decisions[11]. Finally, note that there is not always sufficient data to train and/or test the model properly, therefore it might be difficult to verify the robustness of a model.

(Ill) Performance and inaccuracy

The model will not have a performance of 100%. This in itself (with the current existing technology) is inevitable, and a much heard arguments here is that people make mistakes too. That said, beyond a number of measured model performance, one needs to consider the effect of false positives, a result that indicates that a given condition is present when it is actually absent, and false negatives, a result that indicates the absence of a condition when in reality it is present [11]. The wrong label or other inaccuracies can lead to an outcome where the wrong people/organisations are targeted, without them having done anything. Or vice versa, people/organisations not being outputted by the model, even though they should have been. For instance, using the same example of predicting criminals as in the previous subsection, the search algorithm may seek to identify as many suspects as possible, thus classifying people as criminals unjustly. One could make a choice to decrease the number of false positives here, but that could lead to increase of false negatives, meaning that criminals might possibly go undetected [11]. Such choices then, and the harm and importance we attribute to accurate decisions versus the societal goals desired, need to be weighed carefully.

2.2.2. Socio-technical-hazards

Socio-technical hazards are hazards that exist more broadly for algorithms vis-à-vis their context. Such hazards might sometimes be overlooked, as they are not always easily pinpointed to the AI-system specifically. The following hazards are discussed below. First, the focus on algorithms at the expense of encompassing the larger system. Second, unwanted bias, possibly leading to unfairness and discrimination. Third, the illusion of objectivity of decisions made by or with an algorithm. Fourth, the hazards that are brought by the complexity of the wicked problems tried to be improved by algorithms. Fifth, the opaqueness of systems in which the algorithm works, leading to difficulty in identifying if, when and why errors stem from the algorithm, and when they stem from elsewhere in the system. Sixth, the ambiguity of rules and regulations in a relatively new area of legislation. And seventh, the hazard that responsibilities are not always clear, leading to a so-called 'responsibility gap'.

Algorithmic-centred approach

Coming back to the point made in the introduction of this section 2.2., it may be clear that focusing too much on the algorithmic development neglects their co-existing social factors, including organisational and human factors. Escobar [71] argues that design traditionally focuses technocratic aspects, therefore vouches for an ontological reorientation towards user-centred design, focused on "human experience and life itself" [p. 48]. Positioned within a science technology (S&T) perspective, Chant [72] views 'the social' and 'the technological' to have a dialectic relationship in which "science and technology are recognised as being socially shaped, but, in turn, have due weight in (...) subsequent social change". Converting these words to the research context; algorithmic systems are not only influenced by social factors, but social mechanisms are mutually influenced by the algorithms.

From a critical and feminist perspective, Muller describes a phenomenon where an overly technocratic view on socio-technical decision domains, leads to 'forgetting' in data practices[73]. They argue that within every step of the technical data modeling cycle, reality is somewhat simplified. In that process, information gets lost or forgotten. "The decisions about the definitions of data are quickly forgotten beneath a series of additional decisions, opportunities, improvisations, assumptions and enactments, each of which renders previous human actions less and less known" [73, p. 12]. Marginalised or at-risk groups in society especially are worst off in such forgetting practices, since their voice does not survive the simplification or 'forgetting' process[73]. To acknowledge the added value of humans to data then, and prevent harmful 'forgetting' in data practices, considering algorithms beyond an algorithmic-centred approach is important. Moreover, another potential harm of a single-minded algorithm-centred view, is that it leaves one blinded for potential harms outside of the immediate algorithm's surroundings. An example here is the influence of increased information technologies on the environment, often overseen since it is not an easily detectable one-on-one relation.

Unwanted bias, unfairness and discrimination

Closely related to the quality of data, is the unwanted biased outputs a model can generate. Inputting specific data can lead to inadvertent bias [47]. Human bias such as gender bias and race bias, may be inherited by AI. This is largely because AI models are trained by humans and the data sets inputted are made by humans, therefore the existing biases may be exhibited in the real life application of AI[70]. On top of that, the background, culture and interpretations of both the model's programmers- and users can knowingly and unknowingly introduce and sustain unwanted bias to the model. "Any algorithm will eventually have biases, as it is based on historical data to make predictions about the future, whereas,in the meantime, the situation might have changed" [11, p. 489]. When the hazard of unwanted bias or unfairness is introduced, it might lead to unfair- or discriminatory decisions. Thus, figuring out how to program and train models without human unwanted biases or discriminatory outcomes is critical[70]. As a result, whilst the introduction of unwanted bias in the model is largely technical in nature, both the assumptions made that were fed into the model as well as potential consequences such as unfair decisions or discrimination are socio-technical in nature.

The illusion of objectivity

Forces behind some of the pitfalls of models were already described in 1993, by Bankes[74]. He describes that: "At one extreme, computer technologists will point to inadequacies of existing software tools, and suggest that more advanced technologies such as object-oriented simulation languages or expert systems could provide a fix. On the other extreme, critics of the sue of models point to the difficulty of validating models in the social sciences and the propensity for models to obfuscate as much as they illuminate."[74, p. 427]. Bankes makes a point that models are often presented as the 'objective' or ‘scientific’ truth, whilst they in itself are just one representation of the world[74].

A key premise of the use of AI within public decision-making, is the idea that better information will lead to better decision-making in the subsequent occurrences[32]. This taps into the idea of bureaucratic decision-making as highly organised and systematic. Often, terms like evidence-based decision-making and data-driven decisions are used to refer to this field [33]. These terms suggest that this kind of decision-making is more rational and that data are facts. However, section 2.1.1 already showed that the reaching of public decisions is often not clear or highly organised at all. An unjustified illusion of objectivity then in a field of contingency, can lead to insufficient critical reflection, and, to its extreme to full reliance on a system without ensuring accountability measures.

Complexity of wicked problems

Building on the previous point, inherent to the nature of the topics dealt with by governmental bodies, they are often challenging and operating within an unpredictable socioeconomic environment. Effective decisions must thus be made within this context of unavoidable uncertainty, also labelled as decision-making under deep uncertainty [75]. It does not mean the topics should not be addressed by models at all, rather, it means that the decision-making process should be designed with the utmost care, incorporating as a minimum the origin of the data, the implicit- and explicit assumptions going into the system and the algorithmic governance as key elements to avoid undesired outcomes [11].

Unclarity and novelty of rules and regulation

The introduction chapter of this research already mentioned that the European Commission has proposed an AI Act. Thereby, either in reaction to the proposed Act or independently thereof, many public bodies are drafting frameworks, guidelines, registries and regulations on the topic, a few examples are the European Assessment List for Trustworthy Artificial Intelligence (ALTAI) for self-assessment, guidelines of the Ministry of Justice and Defence, the data aid decision aid (DEDA) framework and the algorithm register of the municipality of Amsterdam [76–79]. However, three points of critique or concern can be discerned from literature here.
Firstly, although ideally legislation should be reflected in algorithms, in reality the legislation is written at a high level of abstraction and thus open to interpretation [43]. The proliferation of different types of guidelines, rules and legislation does not always make it easy to know which rules need to be followed and the level of abstraction does not help in practically knowing how to follow them. Georgieva et al. [80] identify a gap between the ethical frameworks on the one hand, and practically useable operationalisation of ethics in the development of AI-based services or products on the other.
Secondly, as the field of AI develops quickly, one might expect the rules and regulations to keep up. However, the legislative process takes time. Taking the AI Act as an example, the European Commission published a proposal to regulate artificial intelligence in the European Union on the 21st of April 2021 [81]. After many revisions and approvals, the most current official update is given on the 6th of December 2022 (over 1,5 years later), it says that ”the Council of the EU adopted its common position on the AI Act” [81].

Thirdly, taking the High-Level Expert Group on Artificial Intelligence, an influential advisory body to the European Commission on AI and AI Ethics who also wrote the ALTAI, as an example, the composition of the group, strongly representing industry, combined with the lack of proper civil society engagement, invoked discussions on the legitimacy of the production of the guidelines [80, 82–84]. Regarding rules and regulations more generally then, there are concerns about ’ethics washing’, about representation, inclusiveness and negligence of structural background injustices [80, 82].

Responsibility gap

In 2004, Matthias [85] introduced the problem of the ’responsibility gap’ referring to intelligent systems equipped with the ability to learn from the interaction with other agents and the environment will make human control over their behaviour difficult, if not impossible, and yet, human responsibility requires such knowledge and control [85, 86]. Attributing responsibility to persons then, might be difficult or at least altered within the context of artificial intelligence. Matthias poses a dilemma for society here: either decide not to use this kind of machines any more, which he argues is not a realistic option) or face a responsibility gap, which cannot be bridged by traditional concepts of responsibility ascription [85].

Santoni and Mecacci broaden the concept of the responsibility gap as identified by Matthias, and define it as a set of at least four interconnected problems; 1). gaps in culpability, 2). moral accountability, 3). public accountability and 4). active responsibility caused by different sources such as technical, organisational, legal, ethical, and societal sources [86]. Their aim is to be able to better see which kind of responsibility is threatened by which aspect of AI and why this matters. Below, the four types of responsibility gaps are briefly explained.

Culpability gaps refer to the blameworthiness if things go wrong, if important interests or rights are infringed. In the context of AI-based systems, this may convert into a problem, once domain experts base their decisions on an algorithm and they find themselves acting wrongly, yet they could not have predicted the wrong outcome nor reasonably avoided it. The example Santoni gives here, is that of an accident by a self-driving car; if the reasoning scheme underlying the systems' actions is not easily accessible to their controllers, regulators or even their designers, can anyone really be put to blame? [86]. Moral accountability gaps next, refer to instances where persons are called to respond to their choices or actions, less so with the intention to blame them, but more so to understand their moral reasoning. The legal philosopher John Gardner calls it 'basic responsibility', regarding it as a core aspect of what it means to be reflective as a person in society [87]. It may be blurred by the introduction of AI-systems however, since "the capacity for different agents to make sense of the system's decision, might be weakened or sometimes lost, together with their capacity and willingness to engage in meaningful conversation about their role and responsibility that comes with it" [86, p. 1065]. Depending on the context, it may be harmful if individuals lose the capacity to reflect upon their own and the AI system's decision output.

Public accountability gaps next, are pre-eminently relevant and mostly the type of 'accountability' referred to within the context of the public sector. Effective accountability mechanisms may enhance both the effectiveness of complex public decision-making processes as well as its compliance with liberal-democratic values [88]. Somewhat in line with the concerns about legislation not being able to keep-up and lack of legitimacy and representation in the legislative process, there has been a debate if sufficient accountability has been or can be ensured [86, 89]. Many of the discussion point to insufficient transparency or the difficulty to understand for human agents, due to the different, and sometimes inscrutable ways of AI operations. In addition to the technical 'black box' often mentioned, Noto La Diega calls these 'organisational and legal black boxes', created or aggravated by AI in public administration practices [90]. Finally, the active responsibility gap, concerns the goals, values and legal norms professionals such as the AI engineers are supposed to promote and comply with [86]. Active stands as opposed to passive responsibility, where retrospectively moral and legal consequences must be faced when something goes wrong. Where active responsibility is desirable, as it may prevent the need for passive responsibility, it is not always easy. Individual engineers may promote societal goods actively for instance, yet their work is embedded in networks and institutions they do not always have influence over [86]. Moreover, people are not always aware of these moral and social obligations. Or if they are, it is not always clear what the (societal) role and responsibilities is when it comes to the algorithms [86].

In short, the introduction of AI-systems within public sector decision making may pose a few responsibility gaps; it is difficult to comprehend which are the aspects a human is to blame for, it is difficult to rationale the decisions made to oneself or others, if one does not understand the system, transparency issues and organisational and legal black boxes, on top of the earlier introduced algorithmic black box, make public accountability an even more complex phenomenon. Active responsibility finally, though desirable, is not always acknowledged as such, and if it, is not always easy when embedded in and dependent on a larger organisation. One might notice, that these challenges are closely related to the awareness of such responsibility gaps, but also to understanding and explanations of the AI-systems involved. Potentially then, improved explainability, as will be elaborated upon in section 2.3, might help address some of these challenges.

Opacity

The last hazard discussed here, posing challenges to the responsible AI implementation process, is closely intertwined with the idea of explainability, closely looked into in the next section. Burrell [2] differentiates between three forms of opacity for Machine Learning algorithms: 1). opacity as intentional corporate or state secrecy, 2). opacity as technical illiteracy, and 3). opacity that arises from the characteristics of ML algorithms and the scale required to apply them usefully. Each of the above should be addressed differently, to acquire a more luminescent and thereby transparent ML system. Answering to the first form of opacity, one needs to carefully consider if the opacity is really necessary to protect the public good, or if 'hiding' data from the public has different intentions. Some authors are skeptical that the current extent of algorithmic opacity may not be justified and are rather a product of lagging regulation [2]. Burrell proposes responses to this form of opacity, for instance to make the code available for scrutiny, or if that is not possible, to use an independent, 'trusted auditor', who can maintain secrecy while serving the public interest.

Tackling technical illiteracy next, is more complicated, due to the level of technical illiteracy, and potential disinterest or resistance amongst certain 'technically illiterate' audiences [6, 46, 91]. The level of technical illiteracy requires differences in enlightening the algorithms; one could describe the code to non-expert actor-groups and users. Alternatively, one can show simplified or surrogate models of a more complicated ML algorithm. The latter would already require to be more technologically literate, so to say. However, explaining code itself requires some basic level of understanding of the concept of coding. Next, studies show that people are not always interested in the ins- and outs of the workings of a machine [6, 91]. Often, people just need a system to do its job and they assume that others take care of the preconditions. Going to the third form of opacity then, even if we manage to get people interested in the workings of the ML system, multiple authors refer to a 'black-box' when it comes to ML algorithms, since the internal logic of the algorithm is defined by the model and not by human beings [2, 3, 45, 46]. In the words of Burrell: "Algorithms are simple mathematical formulas that nobody understands" [2]. This form of opacity is explained in the following quote: "While datasets may be extremely large but possible to comprehend and code may be written with clarity, the interplay between the two in the mechanism of the algorithm is what yields the complexity" [2].

The challenges surrounding opaqueness could lead to harms in various forms. One of these potential harms, in the context of responsible use of an AI system, is the inability to know why an algorithm produces certain outcomes. It becomes difficult then to pinpoint the impacts of an algorithm, or to identify errors, with the inability to intervene if these outcomes or the rationale behind the outcomes are undesirable as a result.

To summarise, the implementation of AI within the public sector has technical and socio-technical hazards, amongst which at least the quality and quantity of data, potential (ill)performance and inaccuracy, unwanted bias, (too) algorithmic-centred of an approach, the illusion of objectivity, the complexity of wicked problems, unclarity and novelty of rules and regulations, a responsibility gap and the opaqueness of the system. Knowing these technical and socio-technical hazards is important, since they may lead to harms if not considered carefully. The hazards themselves may pose challenges to the AI implementation processes, but so does understanding and accounting for the hazards. One of the ways to accommodate for at least a few of these challenges, and an increasingly studied concept, is Explainable AI. The following section looks into Explainable AI as a way to respond to the AI hazards responsibly. It dives into its definition, methods, audiences, purpose, types and controversies, to have a solid foundation for the empirical study of AI implementation dynamics and the role and possible contribution of XAI in practice.

2.3. Explainable AI

Before looking at the ways Explainable AI are - or can be - used for public decision-making practices, the concept of 'Explainable AI' itself requires some attention. Acknowledging the diversity and complexity of making the actions of machines comprehensible to humans, has correspondingly brought about a dramatic rise of a call for interpretable and Explainable AI [14]. The chapter starts by understanding why Responsible AI, and Explainable AI as one of these 'responsible' values [20], have emerged. Next, the concept itself is looked into. The term XAI is specified both from a computer science as well as a multi-actor perspective. Ensuing, types of explanations are looked at, followed by controversies and trade-offs within explainability.

2.3.1. Responsible and explainable AI as a response to AI hazards

To answer to the concerns and potential harms caused by AI systems, the demand for 'ethical AI', 'responsible AI' or 'trustworthy AI' has found its way into the playing field [8, 92]. The exact operationalisation of these concepts could be a research topic in itself, however, that is not the aim of this study. Therefore, to have a shared understanding of Responsible AI, this research takes what the European Commission defines as trustworthy as a baseline. According to the Commission's ethical guidelines, responsible AI should be: 1). lawful - respecting all applicable laws and regulations, 2). ethical - respecting ethical principles and values and 3). robust - both from a technical perspective while taking into account its social environment [93]. The Commission translated the ethics guidelines to an assessment list for trustworthy artificial intelligence (ALTAI)[76]. The ALTAI defines 7 values for trustworthy AI, namely:

1. Human Agency and Oversight;
2. Technical Robustness and Safety;
3. Privacy and Data Governance;
4. Transparency;
5. Diversity, Non-discrimination and Fairness;
6. Societal and Environmental Well-being;
7. Accountability.

The need for Explainable AI

All of these seven values, and arguably more, are important for the assurance of responsible AI. This study focuses on the value of Explainable AI, which is described within the fourth value of 'transparency' in the ALTAI. Explainability and transparency are closely interlinked, and sometimes even used interchangeably. Where explainability is more closely related to 'understanding', transparency goes beyond that and requires a full ability to 'see through the system'. Within public administration processes, full transparency is often not achievable, for instance because one needs to deal with sensitive data or because certain data is not available [94]. In some cases, people argue full transparency is not even desirable. Subsequently this research focuses primarily on explainability, as crucial and arguably more relevant than transparency for the types of systems studied in this report. Janssen et al. [11] argue that much of the current research into public sector AI is focused on the development and implementation of AI for (semi) automated decision-making, yet such AI applications are criticised for their opaqueness and lack of explanation [2]. The lack of transparency in increasingly complex models, they argue, create doubts about the use of these models. Without understanding them, humans cannot decide if these AI models are socially beneficial, trustworthy, safe, and fair. That said, attaining such explainability is less clear-cut than it looks. The checklist provided by the Commission for instance, entails only two questions for explainability:

1. Did you explain the decision(s) of the AI system to the users?
2. Do you continuously survey the users if they understand the decision(s) of the AI system?

Based on literature and practice, just these two questions are not enough to attain meaningful explanations. They do not indicate what the quality of the explanation is, they do not guide in the type of explanations necessary, and they merely focus on explaining the decision to 'the user', without specifying who the user is or how different users, especially in different contexts, may need different explanations. Diving into the meaning, relevance and types of Explainable AI then, will be the aim of the remainder of the literature Review.

2.3.2. Concept and definition of Explainable AI

To refer to various branches and paradigms of research concerning explanations within the AI realm, eXplainable AI (XAI) is often used as an umbrella term [95]. The number of XAI publications has increased significantly over the last couple of years and especially after 2017 [19, 95]. This resurgence is driven by evidence that many AI applications have limited take up, or are not appropriated at all, due to ethical concerns and a lack of trust on behalf of their users [12]. Multiple drivers have been mentioned for the increasing interest in XAI, including legislative changes (such as the GDPR which went into effect in May 2018 and the European AI Act which is coming soon), increasing investments by industry and governments and growing concern from the general public [96]. Moreover, the running hypothesis is that by building more transparent, interpretable, or explainable systems, users will be better equipped to understand and therefore trust the intelligent agent [12].

Ironically, even though articles on the subject have been abundant, for a concept that is so closely related to understanding it is surprisingly difficult to comprehend what the term means. In literature, Explainable AI often refers to processes where the opacity surrounding AI is (partially) elucidated [2, 6, 45]. 'Explanation' is closely related to understanding and as such is more complex than merely disclosing information. Linking back to the three forms of opacity mentioned by Burrell (section 2.2.2), each of the three forms of opacity should be addressed differently, to acquire a more luminescent system. With the exception of sensitive cases towards specific targeted audiences, the systems studied in this research are generally not meant to be intentionally opaque. Therefore, the explanations necessary are targeted towards the understanding of the technical AI processes on the one hand, and on applying the algorithms in the wider system's context on the other hand. 'Explainable AI' in this report then, concerns "the ability to explain both technical processes of the algorithm-aided system and the reasoning behind the decisions that the system makes".

'Explanation', as defined by the Cambridge dictionary, means: "The details or reasons that someone gives to make something clear or easy to understand" [97]. Even though this is a rather general definition, making the details of or reasons for AI understandable to the various humans involved, is what lies underneath the idea of 'Explainable AI'. However, note that these explanations necessary 'to make something clear or easy to understand' may differ from person to person. Barredo Arrieta et al. [95] incorporate the differences in audiences in their definition of Explainable AI. For Barredo Arrieta et al., Explainable AI means "Given a certain audience, explainability refers to the details and reasons a model gives to make its functioning clear or easy to understand" [95, p. 84]. Incorporating both the earlier expressed idea that Explainable AI should incorporate both technical understanding and non-technical reasoning, as well as Barredo Arrieta et al.'s interpretation accounting for different audiences, the definition of Explainable AI as referred to in this thesis will be:

"Given a certain audience, explainability refers to the details and reasons to explain both the technical processes of the algorithm-aided system, and the reasoning behind the decisions that the system makes."

2.3.3. The purposes of Explainable AI

Explainable AI is a way to understand the model, the system and the system's outcomes, i.e. the decisions made. Thereby, it is a means to know who is responsible for what. Moreover, it is a means to enable the possibility of changing or improving the system or its outcomes, whether by the developer, decision maker or the data subject. Five outlined reasons to view Explainable AI as a prerequisite to use algorithms responsibly, are given below. The reasons are derived from literature concerning the purpose and goals of Explainable AI [11, 14, 95, 98–100]. One might argue to use valid supplementary reasons for the importance of explainability. Given the scope of this research, these are at least considered relevant for public sector decision-practices.

1. **Model-improvement.** First, creating insight in the workings of the algorithms helps make better and more accurate models. Debugging becomes easier, and by understanding how specific outputs are produced, model performance and accuracy can be improved more easily[98].

2. **Impact assessment, control & accountability.** Second, not only should the model be better understood, linking back to algorithm hazards and harms, so should the system. Explainable AI here becomes a prerequisite to understand the rationale behind the algorithm's use and understanding its (potential) impact and vulnerabilities. Correspondingly, explainability for this purpose can improve effective management of those responses [98]. Closely related is that without proper explanations, it would be difficult or maybe even impossible to know who is accountable and for

what. As Janssen et al.(2022) explain: “lack of transparency or explainability hinders accountability and meaningful scrutiny of algorithmic solutions”.

3. **Improved decisions.** Third, in order to make a well-informed decision, the decision makers should understand the limitations and the potential of the AI-output for their decision making [11]. Raghunathan [101] found that the quality of decisions improves if the decision maker understands the problem and the relationships between problem variables. Besides, decision makers' skills and their experience proved important in decision-making [99]. They affect the outcomes; those already familiar with the domain make better decisions than those without any prior experience.[2]

4. **Contestability.** Fourth, explainable AI is useful to justify and provide reasons for the system's output and recommendations [98]. It is a prerequisite to allow for contesting the decisions made based on AI systems. Without understanding how a decision has been formed, in this case a decision based on AI, one cannot determine whether people are wrongfully affected nor can it be shown that discriminatory practices have taken place. Explainable AI in this regard ”embraces the social right to explain decisions to the public” [11, p. 479].

5. **Learning & evaluation.** The fifth reason distinguished, serves as a way to improve and monitor the AI system, its usefulness and consequences. Explainable AI can help discover the opportunities and limits of the AI system, to learn new facts and gain new insights [14].

In short, explainability is a key-prerequisite to use algorithms responsibly in decision-making, therewith mitigating some of the potential harms as discussed in previous section. Beware though that merely increasing explainability does not prevent that incorrect or harmful decisions are being made. What it does, is aid in knowing when, why and how these incorrect or harmful decisions have taken place or could take place. As a result, the opportunity is created to do something with the information, to withdraw the decision or redesign systemic flaws.

2.3.4. XAI in Computer Science

Note that research on XAI mostly stems from the discipline of computer science, referring to methods and techniques to elucidate the complicated often self-learning mechanisms in the application of artificial intelligence technology. Its origin aligns with the wish to contrast with the concept of the “black box” in machine learning, where even the machine learning designers cannot explain why the AI arrived at a specific decision[102]. That said, referring back to the types of models used in public sector decision-making however, recall that a model's level of complexity subsequently requires different levels of technical understanding, where some are considered more of a black box or opaque than others [102].

XAI methods

To contextualise the XAI methods, Barredo Arrieta et al. distinguish inherently transparent models (such as decision trees, k-nearest neighbours, decision trees and rule-base learners) and post-hoc explainability methods [95]. Within post-hoc explainability, a further categorisation is made between model-agnostic and model specific explanations [95]. Model-agnostic techniques are designed to be plugged to any model with the intent to extract additional information. Sometimes, simplification techniques are used to generate proxies that mimic their antecedents, with the purpose of having something tractable with less complexity [95]. Other times, the purpose is to merely visualise the models or to estimate the feature relevance.

It goes beyond the scope of this research to discuss all the methods within XAI in-depth, especially since a comprehensive taxonomy and classification of methods is a whole study field in itself. However, a short explanation of some of the frequently used methods is useful. A comparative table of different methods, including a few examples for each, their strengths and limitations, can be seen in figure 2.1 below, distinguishing between intrinsic (interpretable) methods, visual representations, surrogate models and example-based methods.

[2] Some argue that with improved understanding of an algorithm, trust in the system will also improve, though if that is actually the case and to what extent it is desirable, is part of a larger debate [100].

Interpretable or intrinsic (meaning that the model solves a decision task and provides explanations at the same time) AI is described as machine learning models which are inherently interpretable if the training features are primarily meaningful [103]. The idea is that humans can understand the cause of decisions made by models such as shallow decision trees. Whereas post-hoc explanations are used for the other types of XAI methods, this is not the case for interpretable models [95]. That said, Nassar et al. argue that the most successful learning models nowadays are not interpretable, since they are based on ensembles of voting sub-models, such as random forests [103]. Explainable AI interpretation methods are proposed to deal with these cases, divided into two main approaches; model-agnostic approaches and gradient-based which are specific to neural networks.

Visual methods next, make use of plots to show the importance of features with respect to the model's decisions [103]. Partial dependency plots (PDP) are an example of this, and work by marginalising the ML model output to show the relationship between the predicted outcome and a subset of the features. Other alternatives for such visual Explainable AI representations are the M-Plot, the accumulated local affects (ALE) plot, and t-distributed stochastic neighbour embedding (t-SNE).

Surrogate models aim to find a simple function that best emulates the original function. LIME (Local Interpretable Model-Agnostic Explanations) is a method that finds local surrogate models by minimising a loss function[15]. SHAP (SHapley Additive exPlanations), introduced by Lundberg and Lee, is another popular method providing model-agnostic local estimations of the contribution of each feature ML prediction [17, 95, 104]. It is a method to calculate an additive feature importance score for each prediction with three desirable properties; local accuracy, missingness and consistency [95].

Example-based explanation methods provide historical situations similar to the current situation [105]. Examples of example-based explanations are counterfactual explanations such as DiCE or Anchors. A counterfactual explanation is a statement showing which features would have to be different for a desirable outcome to occur, more specifically looking at the smallest change to the feature values that would flip the outcome of the prediction [103]. They provide a "what-if" explanation, so for instance if one were to receive a loan and is denied, the model would tell you what you would have to do (e.g. earn 10.000 euros more) to receive the loan [18, 106]. In contrast, an anchor explanation is a rule that sufficiently 'anchors' the prediction locally regardless of the other feature values [107]. In other words, with an anchor the prediction is (almost) always the same.

Method Type	**Examples**	**Strengths**	**Limitations**
Intrinsic	Decision tree, regression	Direct explanation	Limited performance
Visual	PDP, ALE, t-SNE	User-friendly plots	Limited scope of features (1 or 2)
Surrogate	LIME, SHAP	Works for complex models	Assumes feature independence
Examples	Counterfactuals, anchors	One-shot explanation	Rashomon effect

Table 2.1: Overview different XAI methods with examples, strengths and limitations [95, 103].

2.3.5. XAI from an interdisciplinary and multi-actor perspective

Despite the promise of assisting human decision making through a AI-driven approach, non-computing professionals often find it challenging to understand how a black box system transforms their initial input into a final decision [51]. To meaningfully address XAI within the context of public decision-making, as per suggested by the definition used in this research, one must go beyond modeling and computer science explanations. To quote Miller [12, p. 2]: "The solution to explainable AI is not just 'more AI'. Ultimately, it is a concept introduced to address a computer-human interaction problem, and its should be operationalised as such."

Where the term within research on computer science is predominantly focused on model explanations, the European Union's explanation of Explainability leans entirely to the other side. Explainability there, is defined as: "Feature of an AI system that is intelligible to non-experts. An AI system is intelligible if its functionality and operations can be explained non technically to a person not skilled in the art"[76, p. 26]. As appealing as this may sound to many public decision-making domain experts, explainability are useful for a technical audience as well. Therefore, coupling a technical and a non-technical side to XAI, one needs to at least identify the following, before knowing which (type of) explanation is most suitable. Firstly: who is the audience? Every audience needs to be provided with different details and addressed with different language. A distinction of explanations for different audiences then, might be the most valuable. Barredo Arrieta et al. [95] distinguish between five key actor-groups for AI systems: 1). Managers and executive board members, 2). Regulatory entities, 3). Expert users of the model, 4). People affected by the model and 5). Data scientists. Even though all defined groups are interesting to study, this research specifically focuses the audience groups one, three and five, as explained earlier and justified in the introduction and methodology (section 1.2 and 3.4). Having said that, note that the people managing, developing and using the algorithmic-supported systems are not homogeneous either. Still, generally trends can be discerned, based on their described roles and responsibilities.

Coupling the different audiences above to the earlier defined purposes for responsible use of AI (section 2.3.3), every audience needs distinct explanations. Even though explanations tailored for every audience might serve different purposes, based on Barredo Arrieta et al. [95] certain audiences would benefit most from certain purposes. XAI for Model-improvement, is most useful for the data scientists. XAI for impact assessment, control and accountability, is most useful for the manager. And XAI for improved decisions, is most useful for the domain experts. XAI for contestability is most useful for the decision subject, who are not researched within the scope of this research. Nevertheless, iterating that an audience might benefit from multiple purposes, contestability is not entirely outside of the scope of this thesis, for instance since the manager or another person accountable for the AI-system will needs to enable inclusion of contestability measures for them to exist in the system at all [11]. Finally, XAI for learning and evaluation are considered most importantly for regulatory entities, if one were to follow Barredo Arrieta strictly. However, even though the (third party) regulatory entities are not studied in this thesis, given the earlier addressed importance of evaluation in all cycles of the AI life-cycle and for all actors (section 2.1.3), the purpose of learning and evaluation is considered important for all actor-groups.

Purpose of XAI	**Actor-group/'Audience'**
1. Model-improvement	Data Scientists
2. Impact assessment, control and accountability	Manager
3. Decision-improvement	Domain expert
4. Contestability	Decision subject
5. Learning and evaluation	All actor-groups

Table 2.2: Overview purposes of XAI per audience [95].

2.4. Types of Explanations

The vast majority of applications of XAI focus on explanations of how individual decisions are reached, on a particular date, through or with the use of the algorithms [96]. Dazely et al. [96] call those 'narrow' explanations. Such explanations are important for humans to understand the workings of the AI mechanisms on which decisions are based, yet they rarely provide information about the context, or the assumptions involved; people's believes and motivations, the hypotheses feeding into the models and their interpretations of values, norms or external cultural expectations. This section aims to discern which variable elements need to be factored in, the information needed for the discussed actor-groups, thereafter discussing which types of explanations exist and might be valuable within the researched context.

So far, it has become clear that XAI is not an unambiguous concept. Various approaches could be chosen, trade-offs need to be made, multiple contextual factors need to be taken into consideration and different actor-groups might need different explanations. To create order in a polysemous field, it is helpful to distinguish types of explanations, ideally corresponding to the different information needs of different actor-groups. The categorisation of 'types' here, could be done in various ways, one of which is the following. The ICO identifies two subcategories of explanations: 1. "process-based explanations" and 2. "Outcome-based explanations" [108]. Process-based explanations talk about the governance of the AI system across its design, implementation and evaluation. Outcome-based explanations tell you what happened in the case of a particular decision.

Next, they distinguish six overarching types, not necessarily coupled with an 'audience' but rather based on different elements that could be unpacked within the system.

i). **Rationale explanation**: the reasons that led to a decision, delivered in an accessible and non-technical way.
ii). **Responsibility explanation**: who is involved in the development, management and implementation of an AI system, and who to contact for a human review of a decision.
iii). **Data explanation**: what data has been used in a particular decision and how.
iv). **Fairness explanation**: steps taken across the design and implementation of an AI system to ensure that the decisions it supports are generally unbiased and fair, and whether or not an individual has been treated equitably.
v)., **Safety and performance explanation**: steps taken across the design and implementation of an AI system to maximise the accuracy, reliability, security and robustness of its decisions and behaviours.
vi). **Impact explanation**: steps taken across the design and implementation of an AI system to consider and monitor the impacts that the use of an AI system and its decisions has or may have on an individual, and on wider society

2.4.1. A demand for information needs

Due to the outlined challenges to attain Explainable AI, the term "meaningful" is sometimes included in practice [109]. Meaningful here, "refers to systems providing explanations that are understandable to individual users" [110, p. 2]. Minh et al. [19] suggest that at least five factors influence which type of explanation is *meaningful* to the receiving individual, namely; 1). The decision, 2). The person, 3). The application, 4). The type of data and 5). The setting. In other words, the type of decision, the targeted actor-group, the implementation of the algorithm, the data used and the context all influence the required or feasible type of explanations.

Ideally then, one would have information about all five elements, to cater the explanations accordingly. Unfortunately, often certain elements are either not thought of, the information is not available or, if the information is present it is hard to find it in a structured manner. Since this research and the main research question aims to unravel the value of XAI from the perspective of various actor-groups, these actors will be the starting point to gather both variable and constant factors for meaningful explanations.

2.5. Controversies and trade-offs Explainable AI

By now, many of the advantages and the often necessity for explainability have been highlighted. Nevertheless, explainability of AI systems can have unwanted consequences too [95]. For one, the explanation can be inappropriate or too complex, thereby failing to achieve its objective. Also, explanations could lead to breach of confidentiality, both for the individual but also for the organisation. Some even argue that one should not look at the principle of explainability at all, but rather focus on accountability, for which explainability could be a mean but is not a goal in itself [111].

A few tensions to keep in mind from literature:

1). **Model complexity vs. Explainability.**
One of the main reasons to incorporate AI within decision-making systems is the complexity of the systems and AI is able to tackle or at least aid which such difficult decision-making. Especially if the problem is more complex, it might be logical or necessary to use more high-end and complex algorithms as well. There seems to be an almost inherent tension then, where the most high-performing AI systems are also the least explainable [94]. Less performing or precise algorithms (e.g. regressions)

are often more explainable [14]. Janssen et al. [11] warn that governments should not jump on the bandwagon too fast in moving from rule based to more complex computational algorithms, as this might reduce explainability to an extent that not all actors might be ready for. Additionally however, they add that experienced persons make better decisions, suggesting that a combination of experience and education might nudge towards responsibly working with algorithms for decision-making. D'Acquisto [112] argues that explainability can be a complex problem since machines are often first geared to formal logic and then (possibly) to ethical or legal principles, whereas it might be wise to incorporate legal and ethical principles from the beginning, to prevent conflicts from occurring later on.

2). **AI explained through simplified algorithms vs. a Matryoshkan dolls analogy.**
If complex models are explained by simplified models, when is the 'simplified' model easy enough to be understood? When can it be considered explainable enough? Even though explaining AI with XAI methods that rely on surrogate (i.e. simplifying) models can be helpful, one needs to beware that not yet another algorithm is introduced to explain the prior algorithm, for instance every time a new developer takes over or the algorithm is not understood well-enough. In other words, in analogy to Matryoshkan dolls, one needs to be careful not to keep simplifying algorithms, every time revealing smaller, yet similarly technical variations of the model. Miller warns against approaching explanations by simply adding more algorithms to the mix [12]. They incorporate social sciences in their research of XAI, arguing that explanations are contrastive, selective, that statistical generalisations are not necessarily useful to people, and explanations are social, meaning they are a transfer of knowledge, presented as part of a conversation or interaction between explainer and explainee [12].

3). **Overconfidence vs. Trust.**
Confidence and trust are debated topics in AI and XAI [9, 11, 12, 14]. On the one hand, trust is important, for the AI systems to be used. On the other hand, overconfidence in AI outputs can lead to de-responsibilisation of human stakeholders or the assumption that AI outputs are correct by default[11, 14]. Determining a level of trust in AI which is necessary to use AI, yet also incorporating the expected and unexpected consequences of trust in a system, is an important yet inconclusive tension and, if not considered carefully in XAI, with potentially harmful effects.

4). **Transparency vs. Privacy and confidentiality**
For explanations to achieve their purpose, a certain level of transparency is needed. That said, since confidentiality is often related to the types of AI systems mentioned in this research, there is a fine line between releasing sufficient information, yet ensuring people and companies individual privacy rights and the organisation's confidentiality agreements. One could argue that, for full transparency, one might want to sacrifice certain types of information. On the other hand, one could say that the AI system should not be put in place at all, if it exceeds privacy boundaries. The latter is part of a wider debate, good to mention here for context but out of the scope of the further research [89]. In analogy with Foucault's Panopticon effect [113], many have critiqued AI systems to create a potentially hostile society where the fear of unknowingly being watched, at all times, may slowly take over the way our society functions. On the other hand, proponents of AI systems argue that technological advancement is necessary and/or will help us to gain insight in complex dilemma's we have so far been unable to disentangle, therefore it is okay to occasionally give up some privacy for the greater good. 'Sacrificing' individual privacy, proportionally to the public good returned, is the sweet spot to find then and has to be weighed considerately. Though ultimately, who gets to decide if, when and how it is permissible to use algorithmic systems goes far beyond the mandate of the researched institutions and actors alone.

To summarise, this research aims to see where XAI can help overcoming AI implementation challenges. Several purposes of XAI have been described, in which XAI can theoretically contribute to a better system. Practically, there is still a theoretical gap in 1). Comprehensively understanding the AI implementation challenges to which XAI would contribute, 2). Knowing which information would be useful or essential, distinguishing between different key-actors involved, 3). Ideally outlining which types of explanations could contribute, whilst taking into consideration the potential controversies or trade-offs of XAI.

3
Methodology

A mixed-method qualitative approach is adopted for the execution of this research. Even though AI itself is a mainly quantitative topic, the implementation of AI is embedded in social and organisational settings, therefore a qualitative approach is useful. The explainability of AI as well as the processes leading towards the successful and sometimes less successful implementation of AI within the public sector, are closely related to understanding. There are multiple ways to reach understanding of mechanisms behind the visible surface. Here, given that various perspectives and interpretations of stakeholders are involved, it is valuable and seems suitable to do in-depth analysis of those visions; building from interviews, documents and observations.

Braun [114] structures a qualitative research according to the following steps; 1). the literature study, design and planning, 2). the ethical approval, 3). the recruitment of participants, 3). the data gathering, 4). the transcription of the data, 5). the data analysis and finally, 6). the writing. These steps were followed for this research. Mainly in that order as well, though note that in practice the steps sometimes merged or ran simultaneously, more concretely; adjustments in the planning had to be made along the way, and the recruitment of new participants went alongside the data collection of already recruited interviewees. The research design in the form of a multiple case study approach, the data gathering techniques - zooming in on interviews as the main source of the empirical findings -, the measures taken in terms of validity and reliability, and the scope and limitations of the research methodology are explained in more detail below.

3.1. Research Design: Multiple Case Study

This research uses a multiple case study as prime medium for data gathering and design. According to Yin, a case study should: concern a contemporary phenomenon within its real-life context and have unclear boundaries between a phenomenon and context [115]. Additionally, the researcher should not influence the events studied. This research fits these criteria. AI is a contemporary phenomenon within the real-life context of the Dutch public sector domain. The phenomenon, namely the use of algorithms, is not directly to be detached from its context, and finally the phenomenon is taking place with or within the researcher and the way that is done is not directly influenced by the researcher. Amplifying the study to multiple cases then, gives the opportunity to see if insights are cross-interpretive. Some first indications of generalisation can be drawn from this, however, it is not the primary aim of the research, nor is comparison between the cases.

3.1.1. Case study selection

This thesis is concerned with public sector decision-making practices in the Netherlands. The Netherlands is chosen since both the universities the degree is obtained, the TU Delft (TUD) and Wageningen University & Research (WUR), are situated in the country. Next, multiple levels of Dutch governmental organisations are currently looking into the use of AI algorithms for their decision-making systems; e.g. national government bodies and municipalities. To narrow down the scope for case selection, one government-level has been chosen, which is the national one. This is mainly because national governmental bodies are pre-eminently struggling with the translation of public values in response to the proposal of the European AI Act, thus looking into questions such as Responsibility and Explainability. The European Union is dealing with such questions as well, but on a more abstract and higher-level. Municipalities tend to differ widely in their approach and incorporation of AI, not in the least because of differing means, public services and size.

Besides government-level, the type of national governmental body needs to be determined. Since the research tries to unravel 'practical' use of AI systems, it seems more logical to look into executive governmental bodies (e.g. the *belastingdienst*) than legislative ones (e.g. ministries). Next, even though stated that the primary aim is not to generalise nor compare, for better positioning and understanding of the empirical results within their context, the cases have similar purposes with the use of an AI-system. The choice has been made to look into decision-support AI systems, thus narrowing down the possible cases. Thereby, in the cases studied, the main aim of the AI-system is to assist in supervisory tasks.

Finally, accessibility is a criterion for case selection; being able to find willing respondents for all three identified types of actors, within the given time frame and in an organisation which is in the process of using AI and looking into XAI. Affiliation with the 'AI Oversight Lab' of TNO (the Dutch research institute) who has partnerships with various governmental bodies, assisted in accessibility to cases.

The three criteria above; 1). A national Dutch public sector organisation, 2). Being an executive body with supervisory function aiming to use decision-support mechanisms, 3). Having access to respondents in all three studied actor-groups, has led to the studying of two cases; *De Inspectie Leefomgeving en Transport* (ILT), meaning 'The Inspectorate of Human Environment and Transportation' in English, and *de Immigratie- en Naturalisatie Dienst* (IND), translated to 'Immigration and Naturalisation Service'. Both cases have politically sensitive workflows however, therefore, besides the anonymisation of the data used for analysis and reporting, this thesis will not disclose any traceable information about the workings of the algorithms nor about the persons involved.

3.2. Data Gathering Techniques

The research methods are a combination of literature review, document analysis, observations and interviews are used. Note that the interviews are the main source of data, since the practical manifestation of theoretical concepts and the perspectives of different actors are considered to be key in this research.

3.2.1. Desk research

Scientific literature review

Literature review (reported in the prior chapter) is done based on critical research and evaluation of research on AI in the public sector, AI Implementation and Adoption, Responsible AI, Explainable AI, XAI Methods, Multi-stakeholder perspectives on XAI, Types of Explanations and XAI controversies and trade-offs. The literature studied is used to inform the semi-structured interview protocol on the one hand, and to interpret and contextualise the findings, analysis and discussion of the research.

Document analysis: guidelines and policy documents

Document analysis is done on some key documents such as European documents (the AI Act, the GDPR) and the national Responsible AI guidelines, since much of the momentum for the topic is initiated by these documents and/or the momentum for the topic has initiated these documents. Additionally, internal research documents and review frameworks of the cases are analysed. Note however, that these documents are confidential and as such, are not a main source of research. Rather, they are considered required knowledge to set-up a well-informed interview protocol and to draft useful analyses and recommendations later on.

3.2.2. Observations

Three observations were held within the course of this research. 1). First, the observation of a meeting where representatives of three actor groups of case A were present. The observation was used to inform the researcher about the case study, as well as understand the way responsible implementation of AI was approached and the challenges there within as perceived by the actors present. 2). Next, a one day field visit observing the daily workflows of a domain expert. Since the respondents interviewed kept referring to the domain experts and the gap in knowledge and understanding between the different actor groups, it seemed almost crucial to observe the way current practice works, and view how AI would or could fit into that, for a better understanding and positioning of the research findings. 3). Lastly, an observation of a meeting where managers & Data Scientists from both case A & B were present and shared similarities and differences on AI implementation practices.

3.2.3. Interviews

Interviews form the main source of insights and empirical results. 16 semi-structured interviews are conducted, divided over two cases and five managers, seven data scientists and four domain experts (overview in table 3.1). Note that one of the two cases is disproportionately represented in terms of numbers of interviewees. However, since there was access to internal documents on the AI implementation process, as well as access to observe a meeting about the responsible implementation of an AI system, where all three key-actor groups were represented, and another meeting with managers and data scientists, the researcher feels they had sufficient understanding to include it as a valuable case study. Additionally, two interviews were conducted from a third national public sector executive body with supervisory function, to verify if the findings could at all be generalised cross-public sector executive bodies. The first of these interviews is used to inform the semi-structured interview protocol, and the last is used to verify the findings from the other interviews. The reason not to include the case as a separate entity, is firstly because merely two respondents were spoken to and secondly, since no background information was available to me as a researcher.

Semi-structured interview protocol

A semi-structured interview is the most common form of holding interviews where ”the researcher has a list of questions but there is scope for the participants to raise issues that the researcher has not anticipated” [114]. A topic list is created to answer to the different sub-research questions, largely based on the findings from literature, in combination with the initial observation and a test-interview held. The general structure of the interview is to ask respondents about AI implementation and XAI perspectives without prescribed definitions, after which more specific questions are asked based on the topic list, though leaving room for further questioning. The topic list can be found in Appendix A.

Interview reporting and analysis

All interviews are transcribed and coded. A transcription here, refers to the written representation of what was said during an interview, with the aim of creating as clear of a complete rendering of what was voiced. Having a literal transcription including slang and ’*erm's*’ avoids interpretation of the interviews before actual analysis is performed. Coding next, refers to the process of identifying aspects of the data that relate to your research question [114]. There are two main approaches in coding of qualitative analysis, one is called ’selective coding’ and the other is called ’complete coding’, where selective coding applies to the pre-identified topics one is interested in, and complete coding refers to anything that is of interest or relevance to answering you research question [114]. For this research, a mix of the two is used. Selective coding is used to report on the general concepts as addressed in the topic list. However, for the concepts that could be related to either implementation challenges or information needs, complete coding is done, with the reasoning that an overview is still missing from literature, therefore any potential answers from the respondents could be of value here.

For the coding, the software ATLAS.ti is used. A range of software programs is available, of which the most commonly used are NVivo and ATLAS.ti [114]. Using such a tool allows for structured organisation of the data, codes and analysis, gives reassurance of comprehensiveness of the coding, and increases efficiency and replicability compared to coding by hand [114]. The codes are used to structure the empirical findings for the later results chapters. It is also a way to stay close to the respondents’ perceptions and answers. And finally, it is a way to easily find how many respondents mentioned certain topics, however note that ”qualitative research is about meaning, not numbers” [114]. In other words, where one might have the tendency to quantitatively know how many respondents mentioned a certain topic, the aim of the research is to generate meaning and merely having less respondents mentioning a topic, does not mean the topic or perception is less valid.

Table overview of interviews and observations

An overview of interviews and observations held, is displayed in the table below.

Method	**Timing**	**Actor**	**Identifier**
Interviews Case A	September 2022-November 2022	Manager	A.1.1
		Manager	A.1.2
		Data Scientist	A.2.1
Interviews Case B	September 2022-January 2023	Manager	B.1.1
		Manager	B.1.2
		Manager	B.1.3
		Data scientist	B.2.1
		Data Scientist	B.2.2
		Data Scientist	B.2.3
		Data Scientist	B.2.4
		Data Scientist	B.2.5
		Data Scientist	B.2.6
		Domain expert	B.3.1
		Domain expert	B.3.2
		Domain expert	B.3.3
		Domain expert	B.3.4
Observations	June 2022	All three actors, case A	C.1
	October 2022	Domain expert, Case B	C.2
	November 2022	Managers and Data Scientists, cases A and B	C.3

Table 3.1: Overview of 16 interviews and 3 observations conducted between September 2022-January 2023, divided over two case studies and three actor-groups.

3.3. Validity and Reliability

To increase the validity and reliability of the research, a few actions have been undertaken. First and foremost, the ethical guidelines of the TU Delft have been followed, and approval of the ethics committee has been granted on the 12th of August 2022. Next, the respondents gave explicit consent to record and analyse the data, given that the data was ethically use as explicated in the TU Delft’s ethical approval procedure. For one of the cases, an additional non-disclosure agreement (NDA) of sensitive information was signed, and the information reported has been checked before publishing.

In total, 16 interviewees with an average of 60 minutes per interview, meant +/- 960 minutes of content worth transcribing. All the transcripts have been anonymised, personal data has been taken out and the reporting within this document happens based on a given identifier, instead of on personal data such as name or title. Moreover, one big advantage of case study research is that multiple data gathering techniques can coexist. This way of gathering data, also known as triangulation, is used to create a holistic picture of the cases described. It increases internal and external validity of the research, because it supports claims by different sources of evidence. In addition to triangulation, the measures already mentioned in subsection 3.2.3 contribute to improved validity and reliability of the data; public insight into the topic list used, literal transcription of interviews, and codes created in ATLAS.ti. A final point worth mentioning is the following. To maintain integrity and understanding as to what is meant by the respondents, the researcher aims to use all the quotes reported in the empirical results within their righteous context.

3.4. Scope and Limitations

The initial scope of the research was only to include the perspective of managers and data scientists. However, after several interviews, every respondent mentioned the challenge of reaching domain experts and grasping their perspective. As a result, it seemed crucial to speak to decision experts as well, and get their perspective on the situation, especially since they are the (potential) users of the AI systems. Thus, the scope of the research widened to include a third actor group. Nevertheless, the fact that it was challenging to find their time for managers and data scientists, forms a challenge to gather respondents for the research as well. As a result, a first limitation is that, comparatively, less respondents within this actor group are interviewed.

Second, ideally an equal amount of respondents for cases A and B would have been interviewed and observed. As mentioned in section 3.2.3, the insights of case A are still considered to be valuable, however, more individual insights from domain experts especially, would have increased the quality of understanding.

Third, to get a full overview of the AI implementation cycle, one would also like to speak to the decision subjects. However, the explicit choice is made to leave actors outside of the respective organisations out of scope, since that could be a whole research in itself.

Fourth, note that the managers interviewed differ in tasks and hierarchical levels. This has the advantage of creating better insight into the organisation and AI system's creation as a whole, but forms a limitation if one would like to compare the managers' perspective. The data scientists as well have different tasks and specialities, but the hierarchical levels tend to be more equally distributed.

Fifth, conducting the interviews in the Dutch language (for the majority of interviews) is a choice, benefit and limitation worth mentioning. The choice is to speak to people in their native language where possible. The benefit is that people express themselves more freely, and since work processes are also done in Dutch, it is easier to communicate their daily practices. The limitation however, is that translation also means a certain degree of interpretation. The choice to translate quotes instead of keep them in the original language, is to avoid traceability to the distinguishable few respondents whose interview was conducted in English (since they expressed to be more proficient in English than in Dutch). As a researcher, I have tried to stay as close to the exact translation of the spoken words as possible, though one can never promise 100% accurateness of interpretation.[1]

Sixth, in qualitative research even more than quantitative research, not only the researched respondents and their world views impact the outcomes of the study, but so does the researcher's. Ideally, one might want to say that research is objective. However, in research dealing with people, the value might exactly be in the acknowledgement that this is not possible nor enough. With a background in public administration, participatory governance as well as in software development (mainly in the Dutch national public sector context), the researcher shares cross-disciplinary ideas with the different respondents, making it easier to talk to- and contextualise the interviews held. At the same time, the research design, the questions asked and the interpretations done, are always subjective to a person's mindset and thinking, making it all the more important and helpful to know the researcher's general departing point of thinking.

Seventh, in an ideal world, factoring in accessibility for case selection and respondents would not be necessary at all, as it forms a limitation in designing the ideal and most comparable research set-up. Realistically however, practice is not easily directed. Therefore, it is a limitation that has been accepted, due to the nature of wanting to study practice, instead of pre-created or proposed measures or hypothetical scenarios.

[1] For those among us who are interested in linguistics, throughout the interviews the Dutch word *uitlegbaarheid* and the English word explainability were used interchangeably.

4

Empirical Findings - AI Implementation Dynamics

This research started from an Explainable AI point of view; what could be the role and possible contribution of the concept within the implementation of public sector AI practices? Nevertheless, throughout the desk research and empirical data gathering, it became clear that AI decision-support mechanisms are hardly reaching implementation at all. For decision-making purposes, rule-based systems derived from manually entered indicators have found their way to systematic use, but learning computer models tend to be in an experimentation phase at best.

Nonetheless, the concept of explainability may still provide insight into how we can overcome certain challenges that AI faces in the public sector. To do so, we must first understand what is at stake in said sector. Interviews were held with three identified main actors in the implementation process; Managers, Data scientists, and Domain experts. This chapter dives into the studied AI implementation practices and aims to disentangle the nature of challenges, currently preventing AI decision-support systems to go through all cycles of the AI life-cycle. It addresses the first sub-research question ”What are practices and challenges of implementing AI in the public sector?”. The challenges will be mapped onto the AI implementation and adoption life-cycle (conceptualised in figure 2.2) to present a comprehensible overview of challenges.

4.1. AI-systems’ Implementation Practices

Following the first section of the literature review, public sector AI-supported decision-making consists of various elements. Moreover, note that every person brings to the world their own view, reciprocally influencing how they interact with- and report on the AI implementation practices. Accordingly, the AI implementation practices in this section constitute of: 1). the respondents’ views on AI (*what* is AI?), 2). the reasons to want to implement AI systems in the studied cases (*why* is AI used?), and 3). the use of AI systems, including the current interaction between the humans and the machine, the humans amongst themselves, and the process of AI implementation (*how* is AI used?). These three elements open up understanding, including to see how the different respondents have different ways of looking at AI, reasons for using AI and preferences in using AI. Understanding these differences and the remaining uncertainties in the AI implementation process now, serve to set the stage for the AI Implementation Challenges in the next section, and to know which challenges can be tackled with Explainable AI, explored in the subsequent chapters.

4.1.1. Respondents’ perspectives on AI

The respondents were asked about their vision on AI, distinguishing between their understanding of the concept of AI, and their attitude towards the subject.

What is AI?

Answering to the question ”What do you view as an AI system?”, the answers differed significantly. Mainly, one could separate the answers into three continuous categories; one where AI is viewed as a broad concept encompassing almost any kind of data. One where only the more complex models are called AI, and business rules or excel sheets are largely regarded ’old-fashioned’ or disregarded at all. And one, though perhaps notably only mentioned by domain experts and not by managers nor data scientists, where people indicate not to really understand the concept. To exemplify the first sentiment: ”Well, I see that very broadly. Artificial Intelligence is the moment that we are processing information with the aim of making new information” (A.1.2). Such a broad definition, where AI is largely related to working ’data-driven’ or ’information-driven’ including digitisation and rule-based systems. Within the second sentiment, the line is often drawn between AI where the model is considered less straightforward, and knowledge-based models, where indicators are manually entered by the data scientists. One of the respondents carefully voiced confusion if the decision-support mechanisms in one of the cases really are considered to be AI at all - ”I keep coming back to the fact that we hardly use actual models” (A.2.1) - arguing that the algorithms used are not considered to be self-learning or automated. To exemplify the third sentiment: ”The data scientists chose AI for the model. And they also chose Machine Learning. I still do not know if those are the same thing, but it is up to the data scientists to develop that part” (B.3.3). Asking what AI means for another domain expert comes the answer: ”My vision about AI is that we do not know a lot of things. I think it is still in the *kinderschoenen* (literally: ’children’s shoes’, figuratively: ’in its infancy/early days’). You can go really far. I believe it works on patterns now, which repeat themselves? But I am not sure” (B.3.4).

The perspectives on ”What is AI?” did not merely differ in *what* respondents answered but also in the *way* the different respondents answered. In every one of the four interviews held with domain experts, the discussions around AI, or XAI for that matter, was never unaccompanied by examples from practice. To give a concrete example; responding to the question ”What are the main reasons for using AI?”, five minutes were spent on giving examples of current ways of working (B.3.2). Anyone who has done qualitative research might think the interview had simply gone off-topic, but the examples were on-topic indeed, working towards the point that innovation had been used in these cases before, and why AI then, could or could not add to such contexts again. AI for the domain experts thus, was considered an accompaniment to those working processes, or not, depending on the respondent, though every time reasoning from their practical workflows. In short, both the data scientists and managers talked about the concept of AI or its context, to then turn to how AI manifests itself in their- and the organisation’s daily practices. The domain experts on the other hand, talked more about their daily practices, to then explore if and how AI could fit there within.

Attitude towards AI

Much alike the current societal debate, attitudes towards the use, righteousness, and desirability were varying. That said, most of the respondents (13 out of 16) were inclined to being positive towards the use of AI, seeing its added value in the organisation already, or alternatively, to see opportunities within the evolving systems. Note a somewhat positive connotation is not strange, since the respondents willing to talk about AI for the sake of research, might tend to be interested in the subject. That said, a generally positive attitude did not mean they were unconditionally favourable. Most respondents indicated some form of caution and/or need for regulation. Data scientists (5/7) mentioned that the technology needs to be embedded in ethical frameworks, and respondents from both manager and domain expert groups went further in voicing their concerns or potential downsides to the implementation of AI. Out of 16, one respondent was notably critical, one respondent was arguably indifferent and one respondent was considerably neutral.[1]

Finally, the respondents expressed to be aware of varying attitudes across the organisation. Words such as ”resistance”, ”lack of trust” on the one hand and ”data-savvy” on the other were used. Moreover, several respondents voiced an observation or hope that time and new employees would influence the general attitude within the organisation. ”Also because we do so many ’trial balloons’. Yeah, AI will definitely be the future” (B.3.4). Three respondents mentioned the soon retirement of a few colleagues

[1] Disclaimer: preventing statements to be traceable to individuals, the respondent identifiers where 1 respondent clearly thinks differently than the others, will not be given.

(B.2.2, B.3.1, B.3.4), in similar lines a respondent voiced: "I know hiring younger new employees is not synonymous with interest in technology, but it does help" (B.2.2).

4.1.2. The reasons for using AI

Why do, should or could AI systems be used within the organisation's current working practices? Those are the questions addressed in this subsection. Efficiency was the most given reason to initiate an AI system. However, beyond efficiency, almost every respondent added that incorporating AI has reasons of improved quality or improved accuracy as well. And finally, two respondents mentioned contributing to a currently unsolved societal problem.

In the words of a respondent "Efficiency is mostly the main reason" (A.1.1). The idea is that a lot of work has to be done, but there are not enough people to do the amount of work. "Ideally, everyone who does not abide the rules, would be investigated, but that is simply not possible because we have limited resources. We presume that with the model then, the chances of finding something increases" (A.1.1). Consequently, especially if tasks are repetitive or information-dense, an algorithm might alleviate some of the tasks and/or make a pre-selection, to target the work more efficiently. Next, improved quality or improved accuracy/a higher rate of correct decision-outcomes were the most-heard reasons to use AI. The following quote sums up the general sentiment well: "We want to move from inspections where they find nothing, to inspections where they find something. So in that sense it is about efficiency, but well, mostly about quality" (B.1.1). A data scientist indicates that AI can offer information the domain expert does not have. "We offer another piece of the reality" (B.2.5). At the same time, they acknowledge not everything can currently be incorporated in the AI system: "Many of the cues to make a decision now, are based on perception and observations. So these are often complex visual and dynamic things (...) Also, domain experts listen to colleagues, so what did their colleagues see last time? We (data scientists) on the other hand, show up with a much broader picture from the data, but the visual, dynamic, complex aspects are not considered in that" (B.2.5). Finally, improved decision-quality is also viewed as avoiding to get into a constricted vision. "Our success rate for decisions is quite high. However, we think there are still some blind spots where we have little to no information. (...) From our experience we always look at the known risks. What do you want to know with AI? We want to know the unknown risks. (...) There are many more things that play their part than we as humans can see at once" (B.3.2). Note however, that there was disagreement if AI broadens one's gaze or not. Another domain expert argued that once the model is trained, it keeps affirming the same types of selections. "That is great, because you can safeguard agreements, but the danger is, that you get into a tunnel vision" (B.3.4).

Finally, the ambition to contribute to a currently unsolved societal problem was explicitly mentioned by two respondents. A domain expert expressed that, once a model is used, it is easier to go to decision subjects before a real trespassing has taken place. "Based on the model, we can visit decision subjects and counsel them. That way, afterwards, at least we need not say 'we did not know' anymore" (B.3.3).

4.1.3. The how of using AI

Having established a background of the perspectives of- and reasons for AI, this sub-section centres around how AI is approached within the cases, as mentioned by the respondents.

Dedicated innovation lab or innovation team

Both cases, and increasingly other governmental organisations as well, have a separate innovation- lab or team, introduced a few years ago and specifically dedicated to introducing innovative data-driven techniques to the organisation. With the existence of such a team, the significance of the subject of AI is detectable, though the exact importance and size of algorithmic decision-support within the team and/or the organisation cannot be read from the mere fact of its existence.

"The idea of our lab is also that we do experiments and show what is possible with data science, with AI" (B.2.2). The respondent continued, "in the beginning, it was building a lot of *modelletjes* (in English this would mean something like 'little models') with machine learning, doing *experimentjes* ('little experiments'). At the start, you see that people say, oh look, nice, and then go on with their daily jobs. At a certain moment and throughout the years however, we are focusing more and more on bringing the experiments to a next level" (B.2.2).

AI as a human assistant

Without being explicitly asked about it, every respondent felt the need to mention that the machine will not replace the human, nor that it will make independent decisions. To give two examples: "The purpose of an AI system is to improve. And to complete. Rather than to replace, we know this", and "Well, I see the machine as an advisor who gives weighty advice, so I really see it as a decision-support system, not as a decision-system" (B2.3; A.1.1). The exact envisioning of how human and machine would interact, was more varied though. Some thought the decision of the model should only be questioned if the result felt odd. Others thought of the outputs of a model as a starting point, which would then completely be handed over to the human to act upon. "I believe it goes hand in hand with craftsmanship (...) in the end, you need *mensenhandjes* ('human hands') to do the labour" (B.3.4). Others still, regarded the outcomes of the model as options one can interact with, for instance seeing the AI model as a "digital person" (A.1.2) who advises, after which a "real person" will decide if they accept the decision or not. Or alternatively, the real person can come back to the decisions later.

The term "tool" was used several times to describe how AI should be used, for instance in the following context: "We need to prevent that people see AI as extra work. It is more a tool, developed to help you (...) so actually to make the data more insightful, but you will still need to do the inspection yourself" (B.1.3).

Human-human interaction

Considering how AI is used, does not just include the interaction between human actors and the AI models, but the interviews reflected it also includes the interaction between humans amongst one another. Following AI-human interaction then, a brief recount of human-human interaction is in place. The frequency of interaction between the developer of a model and other actors differed a lot, not even from case to case, but from algorithm implementation process to algorithm implementation process. Taking an example where the data scientist aimed to arrange frequent meetings with the domain experts, a priori discussing the modeling choices and testing those choices and assumptions along the way. The data scientist showcased a dilemma here, saying that in their experience, it does not work to show a non-functioning model, so how then does one include the non-technical experts optimally in the process? (B.2.1). As a result, and this was indicated in various forms and with various examples, the interaction between humans is yet to be optimised, however in which form and how remains to be experimented with.

The process of AI implementation - in phases

The introduction of this chapter already mentioned that the decision-support systems have not reached deployment, but which phase do they reach, and how? Knowing what happens in the different phases of the AI life-cycle is useful, because the section hereafter arranges the challenges in line with these phases.

Take a look at the following situational picture, to understand the phase of implementation of one of the cases' main AI for decision-support projects. "So we know we have good models, we have tested them, we have met one-to-one with inspectors, and we have started the pilot. However, we haven't had enough feedback from the inspectors to say that the pilot has been, you know, concluded. So we still need some time to let them use the models. If that will happen or not, that's far beyond me. But we are working on that. We're trying to make it as easy for them as possible, but I think it's also just the level we are right now" (B.2.1). Another description of a main project within the other case, goes as follows: "The model was already ready about 1,5 years ago. I know the domain experts were sceptical in the beginning; does it really work? But well, that is just inherent to change. However, I am not really sure why they did not choose to use it at the time. (...) Now however, I know that we will only deploy if we really have permissions of all layers involved" (A.2.1). The main ways of approaching the AI systems, within each of the AI life-cycle phases, is mentioned below.

AI system initiation and pre-conditions phase

Initiation of an algorithmic support-mechanism is done in various ways, and by various actors; there are examples mentioned of members of all the actor groups having initiated an AI system. On the one hand, initiating an algorithmic system is done derived from a question within the organisation; from the business practice (also considered 'bottom-up') or derived from a strategic aspiration (also considered 'top-down') (A.1.1; A.1.2; B.1.3; B2.2; B.3.2). On the other hand, initiation is done starting from the possibilities of the algorithm, or it is simply done to learn about the options and possibilities of the techniques, leaving questions for if and how it will be implemented for later. "Yes, many of the initiatives are started from the innovation lab, but others are also started from programmes" (B.1.3). The programmes, the respondent explains, are a way for the organisation to determine the highest societal challenges and priorities on a yearly basis. These are translated into risk domains, to further consider what to dedicate attention too. Within these domains then, sometimes AI is introduced as a tool (B.1.3).

The pre-conditions described were practical (time, priority) as well as data considerations (sufficiency) and ethical considerations (bias, impact). One of the managers described it as follows: "It always starts with data explorations. If you do not know the data, you should never build an AI system. If you do not know the problem, you should also not start a system nor its application. As soon as you did that; you know the benefit and the data, then you can start recognising a pattern and think if you really want to use a model" (B.1.1).

AI system design and development phase

The choices leading up to the specific design of the algorithm - the choice of the use-case, the model type, the parameters and who will be using it how and when - differ in the different cases. They tend to be a combination of the purpose of the model, feasibility for the data scientist, feasibility within the organisation and "where the energy lies" (B.1.2).

As for types of algorithms developed, regression models, decision trees and random forests, or a combination of the prior are mainly used within the studied context. Note that the specifications of the models are known and mentioned by the data scientists, at least those who built the models but often also by others within the team, and by two of the managers spoken to. One data scientist added the following: "So we use more complicated models such as random forests but only in a use-case that explanations are very simple" (B.2.1). In other words, for decision-support, increasingly machine learning models are used though varying in complexity.

AI system piloting phase

As of yet, the pilot phase has been reached for about two or three projects in both of the organisations researched. Taking one of the pilots as an example, a manager argues: "We are now working with a pilot that has already been running for a while. We also approached the sector and we want to further develop the model. The pilot should be completed soon and then we will be seeing, with the experience we have acquired, well we want to use it in practice. On the one hand that means the decision experts need to substantially use it, on the other hand, we need to ensure that the maintenance of the tool needs to be embedded organisationally" (B.1.3). Since none of the cases discussed an algorithm being in the deployment phase, we will skip the deployment phase here and go straight into the evaluation phase.

Evaluation phase

Evaluation of the algorithms is done within the development teams. Evaluation of the algorithmic systems is also done, generally after the development and often during or after the pilot phase, increasingly with a third-party present, at least in one of the two cases (B.2.6). One of the respondents said "Well, after the development we need to regard the following, right: Is the system really a lot better at filtering than a human being?" (B.3.2).

4.2. AI-system's Implementation Challenges

To summary the above, from the AI implementation practices, we can see that AI tends to be viewed rather broadly by most managers and domain experts, and more oriented towards ML in the eyes of data scientists and the more data savvy managers and domain experts. The respondents tend to be quite positive about the promise of AI, though with reservations on certain ethical aspects and boundaries. Their main reasons for using AI are increasing efficiency, more qualitative or accurate decisions and tackling currently unsolved societal challenges. Where a special innovation team is introduced within the studied cases, and several AI models are technically ready to be used, they are not actually utilised within the wider organisation. This poses a pressing question: *Why?* The answer would have been simple if the models did not work, or if AI is not desirable after all. Fortunately or unfortunately, the truth is more complex. The following quote presents some insight in the complexity at hand: "There's so many things to fix. Building the actual model and generating the explanation is maybe 10% of the time, 15% max. The rest is validating the model with inspectors checking whether they understand the visualisations and making the architecture, the infrastructure behind all the servers and the services. So that they can use it in real-life at their work. It takes time. (B.2.2)". Disentangling the challenges then, impeding to move towards deployment and adoption of the AI systems, is the primary aim of this section.

A combination of factors lies at heart of the implementation challenges in the discussed study domain. A much-heard argument is the novelty of the field of AI. Even though this is an arguably valid reason, it is not a sufficiently insurmountable challenge, since non-public sector domains have been identified to be further ahead with algorithmic decision-making despite a similar level of novelty (reference). Beyond novelty, various challenges have been mentioned by the respondents. Below, an overview of the challenges is visualised in accordance with the AI life-cycle as defined in section 2.1. The challenges are divided into one of three sub-categories, being; 'organisational', 'human', and 'technical' factors. Note that even though the challenges themselves were identified by respondents throughout the interviews, and an indication of context and time within the implementation process was mentioned by them as well, the exact categorisation of the challenges is done by the researcher. Some of the challenges may be placed in two or three of the sub-categories, since there is no hard boundary line. Nevertheless, discerning the challenges per phase in the cycle, and dividing them into types of factors, gives a clearer overview of the interlinked challenges. This chapter does not differentiate between the size of challenges nor determine the level of impediment each of the challenges may or may not pose. Rather, it is meant to be an overview, to make a complex implementation process more insightful, moreover offering a outline as to where and why the challenges might be tackled by XAI in the following chapters.

4.2.1. Overview challenges mapped onto the AI implementation life-cycle

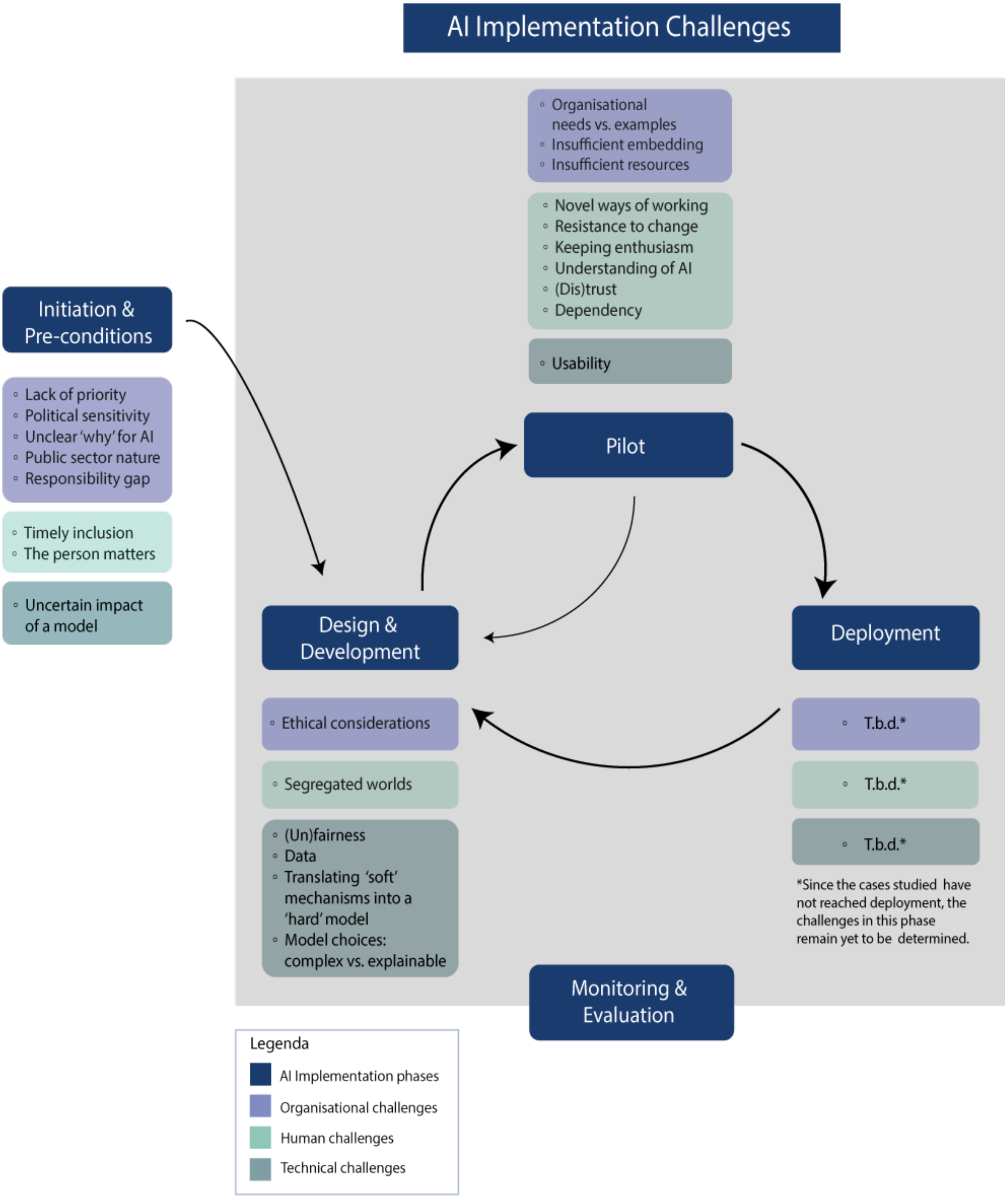

Figure 4.1: Overview AI Implementation Challenges. Mapped per phase of the AI implementation life-cycle (dark blue). In every phase, distinguishing between 'organisational' (purple), 'human' (light green) and 'technical' (dark green) challenges.

4.2.2. Initiation & pre-conditions phase

The challenges discussed here are either related to the initiation of the AI decision-support system, or considered to be pre-conditions, before starting the other phases of implementation.

Priority: managerial, financial and political - *organisational factor*

First and foremost, the lack of a certain ”priority” was a frequently mentioned challenge. The challenge was mentioned by almost every respondent, though in different ways. Time-constraint was mentioned, and often linked to a lack of managerial or financial priority, especially by the domain experts, but also by some data scientists who wished to include the potential users in their development and/or testing process (B.2.1; B.2.2; B.2.5; B.2.6; B.3.1; B.3.2; B.3.3; B.3.4). Already, the domain experts indicated to have too little time to do their work as they deem fit. Dedicating time to either co-developing AI-systems or learning how to use them on top of that, means there is even less time to be allocated for fieldwork. In the words of a respondent: ”In the end, everything comes down to choices” (B.3.4). Managerial priority was mentioned by managers as well, but more from a lack of urgency sensed within the wider organisation (A.1.1; A.1.2; B.1.2; B.1.3). Responding to the experienced lack of priority, one manager reacts: ”I recognise the sentiment. But I think it has more to do with the inexperience with applying such instruments. In the end, it needs to convert into something that is not extra, but just helps within your standard workflows. For that to happen, you need to show little successes at first and take it from there” (B.1.3). Another manager exemplifies the process of (lack of) priority as follows: ”once you agree within an inner circle, there is always an outer circle that starts to interfere, often with the consequence of AI dropping as a priority once again”. Finally, one of the managers compared the topic of AI to climate change, arguing that a few years back, it was still difficult to get people along, but with a wider sense of societal urgency, the organisation picked up on it, in the form of dedicating increased resources as well.

Besides managerial- and financial priority, three respondents mentioned political priority as well. One respondent argued that managers, and the politicians they report to, are judged on their short-term performance. If AI-innovation is predicted to have long-term positive effects, yet these positive effects are uncertain as well as *long-term*, that does not make for a politically viable strategy (B.3.2). Vice versa however, a respondent argued that the European AI act and several AI legislative and public sector organisation frameworks have incentivised political priority for the subject (B.1.3). Thirdly, a respondent mentioned that the ’toeslagenaffaire’ increased political priority for the ethical aspects of implementing AI. ”After things went down with the toeslagenaffaire, it became really important to do it well. As a result, management wants to be comforted” (B.3.3).

Political sensitivity - *organisational factor*

Following political priority, the topics dealt with are politically sensitive as well. The more a topic is politically sensitive, the more difficult it is to initiate and/or implement an AI system, for instance due to the media that jump on top of it, or critical parliament questions. A few statements mentioned in this context were: ”Well, it would have been different if the organisation was not so much in revelation. And if the Deputy Minister says I need to personally have an opinion about the topic, then that’s the way it is” (A.1.1); ”Yes, often the sitting administration is decisive for what gets done at the ministry (...) we are not capable of taking out the political bottom-up” (B.3.4).

Again, the ’toeslagenaffaire’ was mentioned several times in this context, putting the organisations under close observation. ”So yes, you find yourself under a magnifying glass and, where the industry will get away with it, we will not” (B.2.2). Other disrupting societal events or media coverages were mentioned too, such as the ’Vuurwerkramp’ in Enschede (the fireworks disaster, 13th May 2000), a recent news article on the degassing of inland navigation ships (22nd January 2023), and a Zembla episode (9th of September 2019) on the ditching of chemical toxins on the coasts of Asian countries, all of the above to indicate increased societal and political pressure and observance (B.3.2; B.3.4).

At the same time, respondents mention that there are less ’factual’ political factors at play. In the words of one respondent, the fear of being reprimanded or being held accountable for a system you cannot completely oversee, makes people less inclined to participate in such innovation (B.3.2). Another respondent talks about the political questions directed towards the use of algorithms within the context of such a political event, ”Either people think everything is dangerous. And then it becomes very big, very out of proportion. Or AI becomes the solution to all their problems. The truth is of course somewhere in the middle” (B.2.6).

Unclear 'why' of using the AI system - *organisational factor*

Before starting to use an AI system, it seems logical that one knows why the system will be put in place. The general reasons to incorporate AI were mentioned in section 4.1.2. However, when it comes down to specific cases, the 'why' is not always clear to respondents, or the answer is not necessarily shared by the different parties involved. To paraphrase a respondent; we started the project with a question from the domain experts. They were happy with the initial version but then the sentiment altered and the response became; 'okay, nice, we will see when we can use it' (B.2.2). If the goals and expectations are not specifically mentioned at the start then, throughout the implementation process, the reasons might alter or one may lose track of whether or not a goal is attained.

Nature of the public sector - *organisational factor*

A next determined challenge is the nature of the public sector, being relatively slow and relatively unresponsive to (rapid) change. In various contexts, respondents mention the bureaucratic or slow processes, therewith needing patience and time to reach an envisioned change. ”Maybe I oversimplify here, but the biggest problem is not necessarily technical. (...) Of course it is possible, it is not impossible because businesses can also do it. But within the national government, I see many organisations who are still struggling with this step (A.1.1).

Not only bureaucratic processes within the organisation are mentioned, but also the tediousness of regulatory requirements. Three annotations are added by respondents. The first, as mentioned earlier, a slow and bureaucratic nature is not necessarily an impediment to innovative change, since businesses often manage to do so as well. The second, the nature of the public sector is not confined to the implementation of algorithms and can arguably not be changed as such. And the third, slow organisational change was one of the reasons to introduce a dedicated innovation lab or team within the cases, therewith allowing to jump over some of the bureaucratic processes, at least during design and development.

Responsibility gap - *organisational factor*

To counter some of the challenges above, it would help to have well-defined responsibilities; who is responsible and who is accountable for which part of the initiation, development, adoption and evaluation of the AI system? From the interviews however, such roles and responsibilities were by no means clear for all the tasks in the process. Globally speaking, the manager makes the call on initiating an AI system, and is ultimately responsible for unintended outcomes, the data scientists develops the model, and the domain expert uses the model. And yet, in practice, the different responsibilities tend to intermingle. Generally, it is known that a data scientist will develop the model and the decision-maker will not. It becomes more ambiguous for instance, if one looks at who is responsible for including the perspective of the decision-making within the development of the model. Some data scientists mentioned it is part of their job to do so, whereas others indicate its beyond their job or at least not their responsibility.

The following questions especially, 1. Who should have the responsibility to incorporate explainability, or other ethical values within a model for that matter?, 2. Who would be responsible for deployment and maintenance?, and 3. Who should be responsible to make the call on whether or not to go live with a pilot project, raised some reflective and interesting discussions. To take the last question as an example, ”Yes that is a good question indeed. If you leave it to the data scientist who developed the model, they think it is all very nice and they would love to do it. Actually you would want someone else to make the decision, but then really based on understanding the model and the facts, and not just a new article you read in the AD (Dutch newspaper), because that is what happens a lot now” (B.2.6).

How to include actors timely - *human factor*

For an AI-system to work well, a certain synergy between actors is desired, already touched upon in subsection 4.1.3 'human-human interaction'. The managers need to know why and what they say yes to, the data scientists need to understand what is feasible and how to translate AI ideas to practice, the decision makers need to convey how they work now, and where the points of improvement are. And ideally, the decision-subject or other third parties are included in the cycle as well. For all of that to happen, it is important that the actors get timely information.

A frequently mentioned dissatisfaction was that certain actors are included too late in the process, causing unnecessary delays or mismatches in development and adoption later on. ”We develop the algorithms with the users, and with programme managers. However, after that, you see three or four transfer processes. There, often you encounter that not everyone has been involved in the earlier stadiums, or alternatively, even if people are included, someone else who does not agree will come in between and get a say” (B.1.2). ’Timely’ in itself, is not easily defined. Most of the respondents agree that it should be from the beginning; either already during the initiation phase, or in the design and development phase. But whether that should be in the form of providing information or active participation, is less agreed upon. Note that the concepts ’inclusion’, ’involvement’ and ’participation’, the level of each one of them and their most effective form, given the actors involved and the cases researched, could be a whole study in itself. Within the scope of this research however, it is sufficient to know that there is a wish to include the different key-actors from an early stage, yet difficulty in attaining such an aspiration.

The person matters - *human factor*

Of course, the cases studied are doing well in many regards too. Departments are experimenting and pioneering with the initiation, development, usage and embedding of AI-systems, reaching pilot phases at least. Many of the more successful cases have at least one common denominator in the initiation phase; someone who is passionate about models, who is eager to learn, and willing to initiate and sometimes even pull the innovation. Respondents mention that it might be easiest if such a person is an influential technically oriented, person or manager in the organisation, for the innovation to trickle down towards a larger part of the organisation. However, examples were also given of data scientists taking the drive for initiation upon themselves and succeeding therein, and the same goes for decision-makers who seek ways of improvement and turn to algorithms for it, or occasionally an external person who is flown in to initiate such a change. The challenge here however, is that not all the people in the organisation feel affiliated with AI, in other words that the openness and interest towards AI is often person dependent.

Uncertainty in estimating the impact of a model - *organisational & technical factor*

Even if one is very proficient in the workings of AI models, the impact of a model is difficult to pinpoint, given the wicked nature of the problems and the high levels of uncertainty (mentioned in section 2.1.1). Furthermore, the challenge becomes more complex due to the fact that the systems or similar systems are relatively new, therefore there is a limited frame of reference as to which impacts could be considered ’normal’ versus ’dangerous’.

Impact assessment is structurally included in a review framework of one of the cases - with questions such as ”a judgement needs to be made of the (expected) impacts/consequences (and the related risks)”. Even so, the respondents convey phrases like ”The model can have impact that we do not always foresee. We try to include the impact, but it is a difficult question, even more since we do not deal with individuals” (B.2.5) and ”A team lead might not always understand the model well enough to estimate its impact. So I think they need to trust what the inspectors say a lot. And what we say. It is important to include both sides of the story” (B.2.2).

4.2.3. Design & development phase

This subsection looks into the design and development phase of implementation, where the idea molds into algorithms. Again, a distinction is made into challenges from an organisational, human and technical point of view.

Ethical considerations in the modeling process: how to know if third party input is required? -*organisational factor*

In the model development, mostly some sort of procedure is put into place to check the principles and execution of development, confirm if details have not been overlooked or see if improvements can be made. Some respondents talked about the 'four-eye principle' here (where someone else checks your work and therefore you always have two pairs of eyes looking at a piece of code), others about the earlier mentioned review process. Depending on the case and respondents, ethical considerations such as fairness and explainability were also viewed as part of this process. However, a challenge expressed by data scientists here, lies in judging when a call regarding ethical and modeling considerations can be made by themselves, when the four-eye principle or a general discussion amongst team members suffices, when other parties in the organisation need to be involved, or when either the model development or stages later on would benefit from an independent third party expert.

Understanding of the model user's needs -*human factor*

A challenge addressed by almost every data scientist, was that of reaching (potential) users, i.e. the domain experts. Ideally, one would include the domain experts periodically from the start for two reasons; firstly, the model's input, output and outcome would improve because the model is better connected to practice. Secondly, the domain experts will be the ones using the model, therefore many of the implementation or adoption problems later on could already be taken into consideration. The domain expert is often asked for advice on the model input, but even that is not common practice. The following reasons were mentioned. Firstly, in cases where the model is designed from the initiative of the data scientist, they want to see either if the model works first or they did not consider asking such advice from the start. Secondly, the data scientist tried to reach out but did not succeed to speak to the domain experts, at least not structurally. Limited time to do one's own work let alone to add meetings and learning to work with new AI methods are much heard reasons for the difficulty in reaching domain experts. Limited priority from management, and therefore too little allocated time, is also mentioned a few times. Additionally, limited sense of urgency, trust, understanding or knowledge about the subject is heard from quite a few data scientists. Domain experts on the other hand said that it was not just limited time or lack of understanding of a complex matter, but also the sense that even meetings and conversations were much targeted to, and spoken in the jargon of, the data scientists, therefore being experienced to be less useful for the domain expert (B.3.3).

Segregated worlds - *human factor*

In line with the prior challenge, the worlds of the models' design and development and that of using decisions in daily practice often exist completely apart. Once in a while, the 'topics' or actors are brought together in the form of a meeting, discussing practice during model development, or in testing an algorithm in the form of experiments or pilots. Most data scientists could talk for hours about the model development, experiments and the modeling adoption process, whereas domain experts would talk much more about the context. To a large extent this is understandable, as the types of jobs and expertise of a data scientists versus a decision-maker are different in nature. This was even explicitly mentioned a few times, for instance by saying: "I don't know how to do their job, and they don't know how to do mine" (B.2.2). Even though understandable then, the differences in 'language', in ways of speaking to me as a researcher but also to one another, form an expressed challenge for the AI implementation process, exemplified by the following three quotes. "Communication being hard between the different domains, it really impedes. Even when you sit down with someone who has a problem. It's really hard to get that problem in the data science framework. And there are some assumptions made, but those may not be the assumptions the problem comes from", "Yes I notice that even within our team, we sometimes talk another language. For instance if I talk with x, I know both of us want the same thing, but because we come from another background, we don't always understand each other", and "Sometimes we as data scientists think too easily about it. But at the same time others might be too inflexible. It's just another way of working. We have different priorities (...) and that clashes" ((B.2.1; B.2.4; B.2.6).

It does not help here, that the modeling teams and the domain experts are in physically different locations. For one case study, the barrier is even higher since the decision-makers are out in the field a lot, working from a different location in another city. Coming to the office, two hours away, just to sit in a room for just a meeting, feels like an obstacle. Data scientists on the other hand, indicate to have accompanied the decision-makers once or a few times. They say, before Covid19 hit, some data scientists used to work from the domain experts workstation, but now, ”since we only have one set day at the office, we prefer to sit together” (B.2.6).

(Un)fairness - *organisational & technical factor*

”I think, for us it is not a question of: Is it worth it to use it? We know it’s worth it. We have seen the added value. It’s not necessarily a question of how well-performing the model is. Rather, does the model have unwanted bias? That’s what people need to be aware of and be careful with. Is it unfair to a subset of the data? Defined by what we as a society deem sensitive” (B.2.1).

Many ethical concerns and unethical examples have emerged over the last years resulting from unfair model-based decisions. Therefore, considering fairness is important, to prevent unwanted biases on societally sensitive features. The topic was discussed by all respondents in one case and by half of the respondents in the other, though sometimes only after prompting the topic of ethics. Apart from two respondents with a diverging view, indicating that the models are simply objective, ”if it can be objectively indicated that it is a factor of risk (...) well you could call it discrimination, but that is of course not the case, it is just a risk factor” it was a topic considered important by the respondents. That said, most respondents were not too concerned about the unfairness, at least not yet, because of specific initiatives targeted to estimate fairness implications. The idea lives that fairness can and will be ’tested’ after the model has been developed; running fairness tools and algorithms on its workings and outputs. As a result, many of the respondents viewed the topic as something to be considered after development, instead of during development. Additionally, though this was more raised as an addendum and not as a justification for unfair models, the deliberation was raised that now, bias and unfairness also exist in the decisions made by people. ”It has always been the normal way of working, but then within people’s heads. So we know that we need to pay closer attention to certain sector” (A.1.1).

Note that the concepts of discrimination, fairness and bias can be viewed as organisational as well as technical challenges. One respondent has the following point of view on the topics: ”When I say bias and ethics, these are more like high level, say concepts and values and sensitive features. When I say fairness, it’s a technical term for how much this subset of data is being highlighted as risky versus this other subset of data, which is maybe not highlighted as risky. Do we feel that this feature that we use to define the two subsets is a sensitive feature? Is it a problem?” (B.2.1). For other respondents there was less of a strict distinction and the concepts were often used interchangeably, therefore, in this thesis, the topic is considered an organisational challenge in the sense of fairness being a ’high-level’ ethical concept and needs to be embedded as such to prevent potential harmful consequences. Simultaneously, and more mentioned as such in the development phase of the AI life-cycle, it is a technical consideration when it comes to discriminatory and unwantedly biased features and/or outcomes.

Data: sources, insufficiency and historical bias - *organisational & technical factor*

Much like the previous challenge mentioned, the challenge of available, sufficient and qualitative data is an organisational challenge as well as a modeling one. Accurate and sufficient data is crucial for a good working model. Problematically however, such data is not always available. The core challenge here then, is to get access to sufficient and qualitative data to feed the algorithms. A few reasons were mentioned that make access thereto complicated. Firstly, the cooperation between and dependency on other public sector actors for the sharing of the data. Not everyone has access to all sorts of data, and even if access is granted, it might not be possible to verify the data sources as one would be able to if the data was gathered by oneself. Next, one needs sufficient data not just for training the model, but for testing it as well. Besides an arguable insufficiency in data here, COVID19 has caused a gap in available data and/or a distorted picture for the last couple of years. Furthermore, ideally random samples would verify the model’s results, showing that decisions with a model are made with more precision than decisions without using the model. In practice, the data available to train and test a model is already influenced, since the data is based on decisions made by humans for years. Even if one accounts for certain unwanted biases then, the models will remain to have historical bias due to the historically marked data used.

Capturing 'soft' mechanisms in a 'hard' model - *technical factor*

A variation of the following sentence was raised by various respondents: "People do not really run into technology issues. Programming AI is the easiest part I think" (B.2.6). Still, besides fairness and data-related hurdles, a few challenges remain within the modeling domain remain. For one, translating a 'soft' decision mechanism to a 'hard' model poses some challenges. "At the moment domain experts make their decisions based on their own criteria and their experiences. We want to replace that with a risk model (...) but the exact indicators they have? I don't know, that is just the experience of the practitioners" (A.1.1).

Note that discussions are running about the most important features to include here, but also which of the model's outcomes, translated to the real-life setting, would be considered improper. For instance, false positives and false negatives are both undesired outputs. However, the respondents had discussions, and did not always agree on, if is it more harmful to find an organisation who should not have been found, or to not find an organisation who should have been found (C.1). The first might impact the organisation negatively, since they need to do 'extra' work, might experience reputation loss or might lose time and money, the second might be impacting society more, though largely depending on the severity of an infringement not found.

Modeling choices: e.g. complex model versus explainable model - *technical factor*

Making choices about the model is part of the modeler's job but it also has implications for the model; accuracy, explainability for instance, and for its impact; fairness, political implications, privacy. Some data scientists are more aware of these considerations than others, depending on the stage of development and on the amount of interaction with other actors as well. One modeling choice was much mentioned, especially in the context of Explainable AI. In almost all cases, respondents voice to choose a more explainable AI models, decision trees or regression models for example, instead of more complex models such as neural networks. The argumentation behind it, is that the organisation "is not ready for" or "would less easily agree to" more complex and therefore less explainable or understandable models. With these words, though technical options are available, respondents indicated that more advanced AI is still a ballpark away.

4.2.4. Pilot/testing phase

In this subsection, the challenges within the pilot phase are highlighted. Note that some of the challenges here, are also challenges respondents foresee for the deployment phase, though, since those are speculative, the challenges in the deployment phase are left for future research.

Starting from organisational needs versus starting from examples - *organisational factor*

For the pilot to be successful, respondents argue that ideally one would know what the organisational needs are, and build a model to aid these needs. However, often the people understanding and knowing the organisational needs and challenges are not as familiar with the possibilities of AI. If they see an organisational challenge thus, they might not directly think of asking for the development of a model or, if they do, they might be uncertain of the possibilities the model can offer. Hence the creation of an innovation team with, depending on the organisation, team has more or less mandate to initiate projects. Generally it is considered useful for the team to experiment and, with those experiments, create examples for the organisation and also learn if those models could perform well. Nevertheless, starting from the examples, the data scientists mentioned that the model does not always fit the organisational needs, hence the model might not or less easily be accepted. The following discerned paradoxical challenge emerges then: one needs to start from an organisational problem for the model to be helpful, yet one needs examples of a model to know if a model could be helpful.

Insufficient structural embedding of maintenance and ownership - *organisational factor*

A frequently mentioned rather practical challenge, prohibiting a pilot phase to go live, is the organisational question of structural embedding, the ownership and maintenance of the algorithms. In the development phase, the innovation or data science teams have built the models, in the pilot phase, they test them. However, the generally existing idea is that the models need to be maintained and embedded in the more common workflows of the organisation. Besides the innovation teams, both case studies have an IT department, where for instance the documentation systems and 'regular' information systems are accommodated. One manager and one data scientist envision these IT department to take over maintenance. "Well, we really want to use the models in practice, which means the inspectors need to use them, but also that we embed maintenance organisationally (...) That requires making good agreements with the internal IT department" (B.1.3). "Suppose we live in a utopia, where at any moment in time we can think, we used the model for two years now, we learned a lot and now we are going to build a model 2.0 from scratch. I think it is a good idea to do that once in a while, and you have so much knowledge already so, if you have all the data, building a new model is not that complicated. In say five years, that would be together with the analysts of the IT department. And many years ahead, I do not think data science will stay under an innovation lab, but will simply be part of the current information team" (B.2.6).

However, respondents indicate that these departments are too little involved in the current development processes, therefore maintenance remains an unclear challenge. "See, we work in our environment, independently of production. Basically we can develop anything there, we can experiment, make a proof of concept, all those things. The next step is, okay, we have a model, but the moment we want to implement it so that users can really access it, we walk in to a wall, especially if we want to run the models automatically" (A.2.1).

In line with maintenance, questions of ownership are also raised. "Right now, there is a big discussion about ownership in our organisation and it's still undefined. I think what will happen, is that the final ownership will go to the end user. So the data scientist is responsible for the development. We do that to the best of our knowledge, and we also have a review process. Once we deem the product good enough, then we test it. We do a so-called pilot. And then once the pilot is done, it goes into maintenance. (...) They are another team that support the maintenance and they have ownership of the methods that tell you how well the model does: Is it drifting from target?, Does the data change?, Is the model working as it should be? So all of that is on the technical side. But the ownership of the model itself, in an ideal scenario, goes to the end user" (B.2.1).

Resources - *organisational factor*

Having sufficient resources, such as time, money and the necessary infrastructure is ideally tackled in the preconditions phase. However, this is somewhat of a chicken and egg story, since only after development and piloting have shown added value, more resources are allocated, yet sufficient resources are needed to show such added value. Be as it may, the earlier mentioned infrastructure for deployment is still to be sorted and the resource of 'time' has been mentioned by almost every respondent interviewed. One domain expert also mentioned other resources such as lagging internet on the field, but another domain expert reacts they think the internet is fine (B.3.1; B.3.2). When I asked a domain expert "Say we were to set up a a new algorithmic model together, would you be open to that?", they answered "Well, if I look at my own organisation, I recognise that we do not feel included on the one hand, but do not have time for it on the other hand" (B.3.4). So, they go on, "It should be possible if it is supported organisation-wide. (...) But right now, I have my hands full, so it requires choices. This is extra, so if the room is not made, I am also not going to do it. Even though I would like to do it" (B.3.4).

Novel ways of working - *organisational & human factor*

One of the respondents illustrates the new way of working introduced by AI with a metaphor; Going from riding with a horse and carriage to driving in a car, you do not just need the car, you need the people to control it, you need to think about petrol, roads and those are just the simple requirements because people also need rules. And you need roadside assistance. You cannot just run over pedestrians or bikers". In other words, just introducing an algorithm and learning to work with it, requires adjustments beyond the algorithm. Those novel ways of working and their implications become most apparent in the pilot phase.

Some respondents talk about the algorithms as a replacement for some of the tasks done by humans. Others, talk about introducing new workflows, because of the introduction of (predictive) models. Both require a novel way of working, though the organisational and human implications of the latter are bigger. One of the domain experts is very explicit that, in their use-case, a model would be most helpful to move away from an ”ad-hoc” situation and move towards a ”more preventive way of working” (B.3.3). With the rationale that ”preventing is better than curing”, the respondent adds that it still raises challenges, since ”the decision makers are currently not used to working that way”. But also because, in practice, ”we do not always know if we have the jurisdiction to act in this novel fashion” (B.3.3).

Resistance to change - *human factor*

Three main reasons for resistance are mentioned by the respondents: change generally tends to invoke friction with people and within an organisation, some of the changes might be perceived as personal by some of members within the organisation, and the nature of AI itself might cause resistance. Additionally, a few respondents talk about the resistance from the decision subjects. ”It is not as if the decision subject is always happy if we find them. It would have been nice to crawl through the loopholes of the law”.

Exemplifying the first reason for resistance. Interviewee: ”All of a sudden, you need to understand what you haven't understood yet. That is always complicated” (B.1.3). Interviewer: ”Do you think that the domain experts are favourable towards the change?” Interviewee: That depends a lot on the person. We did explicitly use the knowledge of the users in the development of this model. But change is never fun” (B.1.3). Another quote illustrates that change in itself invoked resistance and simply takes time: ”These things ask quite some time. You actually need time to stop what you are doing now and also to try the new for a while. So change simply requires time” (A.1.2).

Next, a few examples of more personally felt resistance: ”Not only is using AI a change from the old system, one needs to beware that some people feel threatened”, ”Suddenly, a domain expert who used to independently make a decision, now gets told by a system that something else may be smarter to do” (A.1.2; B.1.2). A domain expert adds to this point, saying: ”I think it interesting to look at the risk analysis in another light due to the models. But I can also imagine that others feel more resistance, since it also kind of a piece of autonomy you sacrifice” (B.3.2).

Finally, resistance is mentioned where the domain experts especially, but some managers as well are critical or sceptical about the invoked changes by AI. For example, ”Yes, see, the world is changing and things can be made a lot easier. On the other hand, sometimes I wonder, how far do we want to go? Moreover, I do my work with an open mind. Working with information systems means you find more, but it also means you are leaning towards working from a tunnel vision, which is most definitely a disadvantage” (B.3.4). It does not mean that respondents who are critical do not see opportunities. ”It also has to do with expectations” and ”I can also not completely oversee the consequences” are reactions added to the sentiment of resistance.

Enthusiasm: Fostering and/or keeping it alive - *human factor*

From different angles, the level of enthusiasm or motivation to participate and keep participating in the pilots, is mentioned. The data scientists experience resistance and difficulty in finding domain experts to participate, therewith losing enthusiasm for their own projects' implementation. The managers experience difficulty in urging other, less technically inclined, managers to go along with the proposed innovations and/or to convince the practitioners to participate in the implementation cycle. The domain experts, especially those who have participated in a few of the organised sessions, disentangle some of the motivational struggles they experience. One of the respondents says: ”Well, the conversations with different people does not help with the enthusiasm (...) The innovation team works a lot with students like yourself, who then leave again taking with them the knowledge we conveyed. Also, the analysis team works with different tools and knowledge than the innovation team. That there are many conversations with many different people, does not help for the enthusiasm” (B.3.3). Thereby, ”in the pilot phase we experienced that some things are not sorted with the data. The model is only one of the sources we use. If the model would work well, we can fully rely on it. (...) Also, the model worked well for a specific group of decisions, but those fall outside of our legislation, so we cannot do anything with the information” (B.3.3).

Note that one of the data scientists is rather sceptical of the term ’enthusing’. They say: ”If I hear the manager, who is in charge of the trajectory, say ’I will go and enthuse them’, then I think, you shouldn’t be enthusing, you should be guiding what needs to happen”.

Understanding of AI as a general topic - *human factor*

Testing the algorithms in practice, several respondents mention a lack of general knowledge of AI in the field. ”It is related to the question that some people are not as computer savvy” (B.1.2). ”There is also a lack of general knowledge” (A.1.2). ”There are practitioners who literally say, my computer is to sent emails” (B.1.1). Forms of education are mentioned, especially by the managers, to create more general awareness and knowledge. ”You have a national AI course for an AI foundation, but you could of course also include that in the general on-boarding for decision makers” (B.1.1).

(Dis)trust - *human factor*

”So on the one hand, you want people to keep thinking about what the model suggests, and about what do I do with it? On the other hand, it is not necessary to be suspicious” (B.1.2). The balance between the two, is a challenge mentioned by various stakeholders. All domain experts say they would like to see results, to be able to trust the system. Thereby, one of them mentioned ”also, we need the trust the source-data” (B.3.4).

Respondents express the idea that having model outputs that align with the domain experts’ experiences, increases trust in the models (A.1.2; A.2.1; B.2.1; B.2.2; B.2.3; B.2.4; B.3.2; B.3.3). One of the data scientists says, ”Well if the score does not match their gut feeling, they will ignore it” (B.2.5). However, note that some respondents stop after sharing this observation, whilst others go on to be critical of this phenomenon since only trusting a model based on prior experiences leads to bias confirmation. On this topic, a manager talks about the trade-off between trust born from ignorance or fear, and trust born from a healthy critical attitude: ”It is a complicated conversation, because you can only imagine it, if you have experience with it. At the same time, you can only get experience, if you dare to step into the deep. On the one hand it is a lack of imaginative capacity given the specific topic, and on the other hand it is having cold feet. But sometimes it is also right to have a level of distrust” (B.1.2).

Dependency - *human factor*

(Inter)dependency is mentioned with regards to the multitude of ’ketenpartners’[2] and other third party collaborations. ”You have a stream of collaborations with other organisations who do similar investigations” (A.1.1). Dependencies on data then, on third party information systems used, but also simply dependencies working on the ground are mentioned in this regard (C.2; B.3.2; B.3.4).

Also, dependencies amongst actors within the organisation are mentioned (e.g. managers are dependent on other managers, data scientists on domain experts, domain experts of managers, dependencies on the IT department). ”You need everything and everyone, and that makes this game very interesting but also frustrating at times (B.1.2).

Usability - *technical & human factor*

Finally, a challenge mentioned by mainly the potential users of the models, is its usability. Responding to the question: ”Do you feel people will really use the model?”, they answered ”That depends if it is easily applicable. This is one of these things. If it is too complicated or not understandable, people will simply not use it (B.3.2).

4.2.5. Evaluation phase

Evaluation is a topic that ought to be considered structurally in all phases of AI implementation, as considered in the literature review. However, since much of the AI initiation, development, testing, deployment and adoption is still done by means of trial and error, the evaluation is considered and refined on an ’as-we-go’-base. One of the respondents vouches for earlier testing of the model, to then evaluate if it works or not: ”Sometimes we try to go from 0 to 100% at once. Then I think, what if now we

[2] Frequently used term to indicate Dutch public sector partners.

are at 60% and the model helps us to 80%? Then you've already won a lot" (B.3.2). A few respondents emphasise that the pilots will not go towards a 'go-live' situation, before a review has been done (B.1.3; B.2.4; B.2.6). The initiatives put in place to embed evaluation in the review framework- and process are already mentioned. Nevertheless at the same time, the following dilemma is posed: "It is difficult, because beforehand you actually don't have anything. So if you start reviewing before you start, you don't have anything yet, but if you do it after your pilot you might be a little too late. So our ambition, let's put it that way, is that everyone including the data scientists are already involved in the review process, so if they start working on the model, they know which evaluation questions to expect" (B.1.1).

More generally with (policy) initiatives or novel ways of working, often the forefront is heartily discussed and weighted, yet once they have reached the daily conduct of business, things become "as is" and evaluation or (re)development tends to be neglected. Since deployment and structural implementation has not taken place yet, it was not a much mentioned challenge in the studied cases now, but note that evaluation is and will be a recurring matter of importance, and a possible future challenge, for the responsible incorporation of AI systems akin to any systemic change.

4.3. Chapter Summary

This chapter looked into sub-research question 1: "What are practices and challenges of implementing AI in the public sector?". Starting with the practices, the views on AI differed: AI tends to be viewed rather broadly by managers and decision domain experts, sometimes indicating not to know what AI means at all, especially those less affiliated with data in their daily work. AI is viewed more oriented towards ML in the eyes of data scientists and the more data savvy managers and decision domain experts. Though many of the respondents saw a promising role for AI within the public domain, the data scientists and managers more pronounced so than the domain experts, reservations were kept in mind regarding ethical aspects and boundaries of AI. The reasons to introduce AI were partially efficiency-driven, to be able to do the necessary work with the given resources, partially quality-driven, to improve the accuracy of decisions, and partially societally driven, to tackle issues that have so far been difficult to solve. That said, the respondents desired a form of human involvement at all times, making AI a 'tool' or human assistant, rather than a self-operating system. A dedicated innovation- or data team was introduced in both case studies, and multiple AI systems have reached technical development, some currently tested within the organisation. Notwithstanding, even though the algorithms are initiated and developed, many of the AI-systems linger when they are tested in practice. They walk into a figurative wall, where an aggregation of challenges is currently preventing the practices to be successfully accepted by the wider organisation.

> *"Of course, the very first test of the model is; what is the value of the model? Then the next question follows; is the organisation going to be capable of reaping those values from the model in practice? Well, and there, there are a few questions" (B.1.2).*

The second part of this chapter dove into the challenges as experienced throughout the AI implementation life-cycle. Three types of challenges were discerned; organisational, human and technical challenges. The challenges are visually summarised in figure 4.1. In *the initiation and pre-conditions phase*, the organisational factors are most prominent. Lack of (managerial, financial and political) priority, political sensitivity, unclear shared understanding per project of 'why' the system is initiated, the nature of the public sector, and a responsibility gap are identified challenges. Within *the design and development phase* next, the technical challenges are most visible, though arguably many of them are socio-technical rather than strictly technical challenges. Unfairness of the algorithms, sufficiency, quality and historical bias of the data, the translation from a soft decision-making process into hard algorithms, and the choice if a model should be more complex or more explainable, are identified here. In *the pilot phase*, the human factors prevailed most. Introducing novel ways of working, resistance to change, the level of (dis)trust, difficulty to keep enthusiasm alive, a lack of general understanding of AI, and dependency on one another within and outside of the organisation, are identified here.

5

Empirical Findings - XAI in Practice

The second part of empirical findings dives into explainability; addressing sub-research questions 2 and 3 of this thesis: ”What is the current role of XAI in the public sector?” and ”What are the information needs, to improve understanding of the AI system, as defined per stakeholder?”. Following the structure of previous chapter, this chapter starts by considering the perspectives on XAI, the reasons to use XAI, and how XAI is currently approached in the case studies. Next, the chapter aims to unfold the information needs as mentioned per actor-group. These information needs give a first indication of which information would be desired by the respondents, whether or not through XAI specifically. Knowing both the role of XAI as well as the needs from XAI, will inform the opportunities XAI can offer, discussed in the next chapter.

5.1. The Role of XAI

Tthe role of XAI here is constituted of the respondents' perspectives on the topic, the reasons mentioned for the use of XAI, and the ways XAI is applied within the studied cases, i.e. the *What?*, the *Why?* and the *How?* of XAI in practice.

Where the two cases are fairly comparable in their challenges, though nuances between the respective contexts are appropriate, Explainable AI brings a clear distinction. One of the two cases does not use XAI at all, or at least not consciously. However, not explicitly incorporating explainability does not mean the actors do not have ideas or an intuitive notion of the concept. Therefore, the clear distinction between cases will not be made for the XAI perspectives nor for the information needs, discussed in the section hereafter, since intuitive and practical considerations are valuable there. For the reasons to use XAI and the how of using XAI however, the difference is considered both notable and relevant, and will be distinguished as such.

5.1.1. Respondents' perspectives on XAI

Since this thesis understands XAI from a multi-actor perspective, the section will discuss the perspectives per studied actor-group: the managers, data scientists and domain experts.

Managers

To understand the perspectives of managers on XAI, the following quote, responding to the question ”What is explainability in your eyes?”, is an illustrative start: ”If I talk about explainability, it means I just want to see the black box, so to speak. In the algorithms we are developing, but also just in the craftsmanship of the human. Including for instance the bias people have themselves. How can we make such mechanisms transparent?” (A.1.2). Similarly, another manager says: ”Explainability to me is being explainable and reproducible” (B.1.1). They continue that, within the context of explainability, one needs to talk about 'if' and 'how' one has held a conversation with the domain experts. Moreover, as a public sector organisation, one needs to explain how they have had a conversation with the general public as well. ”So the moment one makes a decision based on a model, one needs to be able to explain why they did that" (B.1.1).

Generally, there is a form of consensus amongst the managers that, the more you get towards the 'actual' work on the ground, the more people would benefit from explainability. Unless something seems to be wrong, then other managers higher up in the organisational hierarchy might start to ask critical questions as well. "The moment that the model reaches the *werkvloer* (workplace), and you go to the level of domain experts, it is important that they recognise their way of working in the model (...) That goes beyond reading or applying a dashboard" (B.1.2). Additionally, there is a notion where explainability needs to specifically incorporate explanations beyond the model, and beyond technical terms. "I see explainability as a balancing act between rationality and emotion (...) Moreover, explanations might need to go beyond explaining the specific model. For instance, you might need to explain what 'chance' means in statistics - for instance that a hit rate of 30% does not necessary mean there is a 1 in 3 hit" (B.1.2). Explainable AI from the managers' experience is a very rational approach now, yet much of the trust that needs to grow, or the fear, that needs to be listened to, which they would want to include in Explainable AI, are emotions. Just giving the facts then, will not work to take away the doubts or fears. Therefore, perhaps through conversation though without knowing how exactly, the managers spoken to indicate that there might be something to addressing explainability more intuitively.

Data scientists

Most of the data scientists either have an idea of the theoretical concept of XAI or are properly schooled in- and knowledgeable about XAI, some even with a specific focus on exploring novel (often algorithmic-inclined) techniques. The somewhat simplified average view on explainability then, is that their notion of the concept is in line with XAI, as explained in the computer science field of literature. Responding to the question "What is XAI?", one interviewee sums up various discussions on XAI: "You have different definitions depending on the papers you check. So there's really no consensus. It also depends on explainable to whom; who is this algorithm supposed to be for, and who has to understand it? Is an expert in AI sufficient? Is it sufficient for person A to understand the algorithm, and then convey the information to person B, which is the end user? Or should the algorithm be explainable to the end user? And then it's all how do you quantify explainability as well. (...) Unfortunately it is still a buzzword for a lot of people, which kind of hinders the specificity of the word" (B.2.3).

In their given definitions, the data scientists focus on explanations of the model mostly, with sometimes the extension of including reasoning behind the model, or with specific mentioning of feature importance. For example: "Explainability can be two things, either a model works in a way that it is easily explainable, take the example of a decision tree. Or, you have a surrogate model, something that says something about a complex model you know a so-called black-box model", "So for me, explainability would come down to the point that a practitioner of data science tries to understand what the borderline cases are for which the model doesn't really work that well", or "We have random forests. We do have them, but we try to simplify it maybe to rules or to some few criteria in order to give the user a taste of it, and then they can understand it" (B.2.2; B.2.3; B.2.1). Following up on such a stance, a data scientist adds, "Once we understand the model, the next step is to understand *why* the features contribute to a higher score. That is not yet structurally included in the process. Then you will end up looking at the domain knowledge" (A.2.1). If such domain knowledge is included in the term 'Explainable AI' differs per data scientists.

Domain experts

The theoretical concept as such is not known to the majority of the domain experts (3/4). That said, intuitively they have an idea of what it means or what they think it should mean. "Of course you need to explain: 'This is how we are going to do it. This is what comes out of the model'" (B.3.2). At the same time, the respondent indicated that the context of the work requires knowing the outcomes of the model ahead of time, and that time may not always allow for explanations, if decisions need to be made rapidly. If time is short, they said "how the output is generated exactly, you know, as long as thorough research has been done, it is fine" (B.3.2).

XAI: for whom?

Keeping the audience-dependent nature of XAI in mind, mentioned in the definition of XAI in the literature review of this thesis, explainability is almost always mentioned in the context of the domain expert, including by managers and data scientists. Only after prompting the data scientists if XAI also includes explanations for data scientists, they either said ”of course”, arguing that since it is self-evident it did not come up yet (B.2.4; B.2.5; B.2.6), or that has not been relevant in certain discussed AI-systems, since some of the algorithms are not considered difficult to understand. Here as well, after further questioning, many of the data scientists indicated to see value in Explainable AI for themselves. Two quotes exemplifying that are: ”I also use explainability, but I focus my decision on the model and on the output, whereas the domain expert would focus on the output”, ”Explainability is for the domain experts. But explainability is of course also, or at least the metrics, for data scientists. I think we just agree about that and therefore we do not talk about it” (B.2.3; B.2.4). One of the data scientists reasoned as follows: ”Explainable AI really means different things when you refer to different people. I always view them in two big categories. One category is the experts, one the non-experts. The experts try to really get relationships between the data, relative to the model’s output. So how is each feature or a bundle of features, used to create this one particular score? (...) For the non-experts, it is mainly about giving them sufficient information, and I emphasise the word sufficient, to allow them to use the model properly” (B.2.1).

If Explainable AI also includes specific explanations for managers, in the eyes of data scientists, remains unclear. Regarding the question if managers have enough understanding of models, variations of the following answers were given: ”No I do not think the team leaders know enough about the model itself to make choices. I think they need to trust the domain experts, but also trust what we tell them. I think you need to rely on both sides of the stories” (B.2.2). For the managers as well, initially, Explainable AI was mentioned to explain the model to- and to reach the domain experts. Additionally, when asked if it is also considered for managers themselves, they answered that explainability is particularly helpful for reviewing. Though here, they warned that Explainable AI in a reviewing process, should not just become checking a few questions off the list, but really understanding what the model means (B.1.1; B.1.2).

5.1.2. The purposes for using XAI in practice

To appreciate or see reasons to use XAI, one does not need to have Explainable AI in place, however, in the case explicitly aiming to incorporate XAI, the respondents spent more time thinking about the subject. Therefore, the reasons for using XAI, especially from a data scientist and domain decision expert point of view, will be from case B predominantly. A summary of the purposes mentioned by the respective actor-groups, is given in section 5.1.2.

Managers

Purposes of XAI mentioned for managers themselves

Starting with the managers from the case where XAI is not currently used, one manager answered that Explainable AI for managers would be more so about *following the right steps in the processes*, than really knowing the details or reasoning behind the models: ”Managers do not need to know in detail how it all works. Just knowing if the process has been gone through and if it has gone over all the right tables, if the right people have looked at it, is enough” (A.1.1). However, they added that it is also just all new terrain to them, as a result being unsure how it would be useful in the future. The other manager within the same case attributed larger importance to explainability, mostly to be accountable within their own role, and to increase understanding and use for the role of the domain experts. ”You need to be consistent and also explainable, and know when you have used the model in which capacity. I think that is still a big step we need to make. But I do think it is important, especially if something goes wrong. You need to know why it went wrong and you need to explain it” (A.1.2).

The managers within case B saw a purpose for XAI to help in *accountability*, to create a similar vocabulary and *to explain the model amongst other managers*, and *to communicate externally* whether to decision subjects or to the larger public. ”Well, I do not need to see the details, but I need to be able to explain the model on a higher level. That has to do with accountability, who is responsible for the model in the end? (B.1.3). Moreover, they said it enables them to *estimate or act upon potential* *mistakes made* in the process including AI. ”Even if I make mistakes, I need to at least be able to explain why I did it” (B.1.1). Improved understanding of bias is mentioned, both concerning *insights into the biased assumptions of people*, as well as *understanding the data* and their (historical) bias. ”One should be able to answer the question ”Which data do we use?” in the review process”, and ”Once again, it is important to be clear about your data source, you simply have to. (...) Sometimes there is not enough information, then you also need to be very honest. Factually, if you use an AI model or an Excel sheet, in the end you need to determine what are the sources and how you come to a certain decision. Whether you use a smart piece of software or figure it out yourself, you need to simply explain it” (B.1.3). Finally, sharing explanations with one another was mentioned with the purpose of *creating a critical and open development process*, for everyone to be on the same page and to discuss the model and its assumptions accordingly. ”Well explainability plays a large role in the review process for me (...) having an open and critical attitude, but also with domain experts and people from outside is important. But regarding the quality of our explanations, we might be too much inward-looking” (B.1.2).

Purposes of XAI mentioned for domain experts

Subsection 5.1.1 already mentioned that managers spoke about XAI being useful for domain experts, before talking about purposes for their own work. ”Explanations are important for a domain expert to recognise themselves in the model. Therefore, you need to take them along in the process, show them how the model works and that they can have faith in it” (B.1.2). Moreover, ”maybe a reason of explainability is also just to show how the product works” (B.1.2). The purposes for domain experts, as mentioned by managers, were: To *To take away current tension* caused by insufficient knowledge or awareness, to *recognise their ways of working in the model*, to *how to work with the model* and show how the model works, and to *to trust the model*.

Debates on purposes of XAI - managers

That said, the managers had some doubts if Explainable AI will always help to create understanding. Two quotes exemplify that best: ”Maybe the focus should not be on explainability, but more so on the realisation of what the data can do and conversation about that. (....) The interesting part is, yes we think it should be explainable, but domain experts are not always receptive for it” (B.1.1). Also, ”there is a certain professional expertise the domain expert has and, in explainability, you need to beware that you do not show them how to do their job” (B.1.3).

Data scientists

Purposes of XAI mentioned for data scientists themselves

Within the data scientists as well, a difference in importance could be noticed between case A and B. In case A, explainability was only mentioned in the context of *clearer understanding of responsibilities*. They mentioned that ”the manager needs to manage, therefore to understand the process, the data scientist needs to understand the content and the user needs to trust that the output is correct” (A.2.1). However, the need for XAI methods, according to the data scientist, were not necessary (yet), with the reasoning that the models used are explainable in itself.

In case B, the importance of XAI was acknowledged throughout the team, the team even has one data scientist with the specific responsibility to include explainability within the development process, showing that explainability is on their minds. That said, some data scientists were more sceptical of the impact and use of XAI than others. The main purposes were to gain insights into the more complex models, to improve and *fine-tune the model*, to *verify the model's assumptions* and to increase *(technical) understanding of the unwanted biases of the model*. ”I think XAI is quite a good way for us data scientists to get a better understanding of what the model really does (...) and to really weigh thresholds” (B.2.5).

Purposes of XAI (not) mentioned for managers

Talking about explainability towards the managers was not done at all by the data scientists. Responding to this observation by the interviewer ”What I still miss, is how explainability could help towards management”, a respondents answers ”That is a good point, I did not think about that yet (...) But well, I think they would deal with similar questions as the domain experts” (B.2.4). However, if Explainable AI could help in gaining improved understanding about the model and how, was a question the majority of data scientists did not have an answer to. One of the things that came up during the conversation about the topic, was that you might somehow want to show trade-offs to the managers. ”You might want to show, if we would have programmed it more towards this, than you would have gotten more of those types of outcomes” (B.2.1).

Purposes of XAI mentioned for domain experts

Here as well, 5 out of 6 data scientists talked about explanations being used for the domain experts. Reasons for Explainable AI towards the domain expert were: *enabling better-informed decisions*, *increased use of the models*, *increased trust*. ”You enable a user to know when they should or should not follow the model” (B.2.6). ”Yeah, so it’s maybe not about telling them how the model works, but just giving them sufficient information about the model output that they can decide well, I am able to use this information, or rather not (B.2.1) However, note that the data scientists also indicated to struggle with making the algorithms explainable to the domain experts. One data scientists says: ”After explanation, the domain experts understood the visualisation, but they found it difficult to understand how the model comes to a prediction still” (B.2.2).

Debates on purposes of XAI - data scientists

Nevertheless, a few points of attention are given by the data scientists. 1). First, a difference in model and perception does not directly lead to a change of the model, since respondents indicate mutual bias in people as well. ”If we or the user did not expect something to come out of the model, we will investigate it. For instance, maybe it is the data we have inputted in the model? But we need to realise that the user can also be biased in their assumptions. Therefore, if the features do not align with their worldview, that is absolutely not a direct sign to change the model”(B.2.2). 2). Second, note that not all data scientists think that XAI is helpful in the way it is used now, both for themselves as well as for the domain experts. 3). There was a debate if explaining algorithms to the decision subject is a reason to use Explainable AI. For the domain experts, a data scientist wonders if explanations do not simply increase their confirmation bias, and if the way they are explaining now is really helping to increase understanding: ”Sometimes I think that the more information we give, the more confirmation bias exists. (...) Also with the SHAP values, I wonder if the domain experts really understand it. Some do, don’t get me wrong, but all of them are expected to, and that is the problem. (...) A normal person doesn’t know the difference between a negative and a positive correlation for instance, in the SHAP scores (...) So for the domain experts, we need to find another way to visualise it, or we need to throw a workshop at it, or educate them. I am sceptical about the the we use of data science techniques for domain experts” (B.2.4). For the decision subject next, there was a question of whether they should be given explanations and in which level of detail. ”Should we always explain that we used ML? Yes we should, there will be legislation for that as well. But should we also explain how the model works? I don’t think so. And who should explain that then? Should we, or should the domain expert?” (B.2.4).

Domain experts

For case A, no domain experts were interviewed and in the meeting observed where the domain experts were present, explainability was only mentioned briefly. In this brief interaction, the following reason was mentioned: *understand the assumptions and criteria used to build the model*, in order to understand if the model matched their line of thinking (C.1). Within case B, as pointed to before, the majority of domain experts were not familiar with the term of Explainable AI itself, but they had thought about the importance of explainability. They mentioned the following purposes. The first one was *demonstrate that the model can be trusted*: ”Ah, yes explainability could contribute to trust, if you can demonstrate that it works” (B.3.2). Also, *understanding the input* was regarded an important reason to use explainability (B.3.4). Another reason to use explainability, is to create *the ability to explain the algorithm to the decision subject*.

Debates on purposes of XAI - domain experts

Understanding the details of the model was considered much less important or necessary (B.3.1; B.3.2; B.3.4). Moreover, a domain expert adds, that it is important *how* the understanding or explanations are given, not just that there is any type of explainability (B.3.3). ”There is just really a mismatch between the development of models and practice. So if you talk about explainability, there should really be more time and attention to make the actual translation to practice (...) The data scientists are not aware of the sensitivities in this world nor what the consequences are. So you really need to think about the message you bring and something you cannot say certain things, or need to say them in other words" (B.3.3). However, similar to the questions raised by the data scientist about Explainable AI for decision subjects, a domain expert experiences a tension here, where in theory they want to explain why they make a certain decision, their legislative task and discretionary power now do not necessitate them to do so (B.3.3). So with the introduction of an ML model, a new dilemma is created as to what one needs to explain to a decision subject, due to the model, and what not.

Though 4 respondents mentioned the decision subjects in some form, none of the respondents mentioned contestability. The question, if contestability is or would be considered within the cases' AI-systems, but only two respondents were familiar with the term in the first place, and the others acknowledged that it was an interesting question but they were not aware of measures put into place to account for contestability. The respondents argued that it is not necessary for them to know the model works or what its main features are. What they want to know however, is that it works. "Do data scientists need to explain how the model works? No. They need to be able to show results" (B.3.2). Similarly, another respondent argued: "No, I do not necessarily need to understand the model. I need a useful tool, but whether I understand it or not, you know, I also do not understand my car". However, digging deeper, the interviewee added that: "I do not need to understand the technology behind it, but I do need to understand what the model is fed, it's input, because I need to interpret why I know something" (B.3.4).

Overview purposes for XAI

Since quite a number of reasons have been mentioned throughout the section, a quick tabular overview is given below. Per actor-group, the purposes of XAI as mentioned for themselves, as well as for other actor-groups are distinguished.

Mentioned by which actor	**For whom**	**Purposes of XAI**
Manager	For themselves	- Follow the right steps in the process - To be accountable - Know the model's bias - Know the biased assumptions of people - Communicate to decision subjects and the larger public - Estimate or act upon potential mistakes made - Explain the model towards other managers - Gaining insight into the data used - Creating a critical and open process
	For domain experts	- Take away current tension due to the 'unknown' - Recognise the domain experts' ways of working in the model - Trust in the model - Show how to work with the model
Data Scientist	For themselves	- Insight into more complex models - Fine-tune the model - Verify model's assumptions - (Technical) understanding of unwanted biases of the model - Clearer understanding of responsibilities
	For domain experts	- Enable better-informed decisions - Increased use of models - Increased trust
Domain expert	For themselves	- Understand assumptions and criteria of the model - Demonstrate that the model can be trusted - Understanding the model's input - Be able to explain to the decision subjects

Table 5.1: Reasons to use XAI, as mentioned per actor-group. Differentiated by whom they value the explanations for.

5.1.3. The how of using XAI

This subsection discusses how XAI is used, but also which points of debate are mentioned concerning the use of XAI. These insights showcase that translating the XAI purposes to actual explanations and tools might not always be easy. At the same time, they help to inform and improve the future operationalisation of XAI in practice. Already, it was mentioned that one of the cases does not explicitly use explainability methods, ”We don’t go further than a model that is actually rather easily set-up. (...) Right now, the organisation is not ready to move towards a more complicated model, and to apply such a model explainable” (A.1.2). Therefore, for the remainder of this subsection, the current use of XAI is only deliberating upon for the case where Explainable AI is explicitly experimented with.

Having posed the same question, ”To what extent Explainable AI is currently used?”, three categories of answers came up. The first is through the use of SHAP values (and a few other Explainable AI methods, though these are only mentioned by two respondents), the second is through the development of a Dashboard (based on the SHAP values/feature importance) and the third is within the context of a Review Framework. The XAI methods, dashboard and review framework are discussed below.

XAI methods: mainly SHAP

SHAP, or more generally speaking, feature importance, is used as a tool to increase explainability within the organisation. One of the data scientists also mentioned other methods they use such as Lime, anchor points, and partial dependency plots. Other data scientists are aware of these options, but either indicated to be less familiar with the methods, or to see less value in them, especially if they are still working on getting the models themselves to work, or on other challenges around implementing AI. Respondents indicate that SHAP is chosen, because it is a state of the art tool: ”it was considered state of the art and then you tend to go along. Perhaps if we invest more time and keep researching other options, those might be better. But well, one needs to make choices and it is also a question of how to spend one’s time” (B.2.2).

Debates on the XAI methods

Two master thesis students have studied Explainable AI methods, in the form of SHAP-explanations, within the case study. Haas [116] looked into types of visualisations of SHAP values, reviewing their understandability and usability. The explanations proved suitable for data scientists, but less so for the domain experts. This is in line with the findings from the interviews. A data scientist explained it in the following words: ”We think that it is all very clear and useful for the domain experts. For instance, all of us thought; a scatter plot, of course we should include that in our explanations, it is enlightening. But the domain experts really did not have a clue. We think; it is so simple, even a correlation, we think it is the easiest thing there is, but if you have never learned something like it, than we need to think of something else to visualise or explain it” (B.2.4). Two out of four domain experts, were aware of the existence of a dashboard, addressed in the next subsection, but the terms ’SHAP’ or ’Feature importance’ were not mentioned in the context of explainability. One domain expert says, ”well the XAI methods are sufficient now, in the pilot phase, since there is a close cooperation between a select group of interested data scientists and domain experts. If there are questions, we can quickly ask them, for instance where does this output come from? (...) However, once in production, the connections are less easy, and then XAI methods, as used now, is not enough” (B.3.3).

Note that the question, ”If any XAI methods or type of explainability are used for managers? ” elicited an unanimous ”no”. Multiple respondents indicated that towards the manager, that type of detail level is not required (A.1.1, A.1.2, B.1.1, B.2.1, B.2.2).

Dashboard

To translate the SHAP values to more insightful explanations for the domain experts, a dashboard is designed. The insights as provided by the MSc studies of Haas and Treur, were used to prototype a user-oriented dashboard. A recommendation given by Haas [116] was that using text in combination with visualisations would maximise the usability of SHAP values. Another recommendation was for SHAP experts to look into other ways of intuitively presenting local SHAP-based results. Treur [117] used the insights to develop a dashboard arguing that ”as the inspectors often lack data science skills, it is expected that they have difficulties understanding the detailed and technical visualisations" [117]. They chose for a dashboard, saying that 'generally, designers do not settle for a single visualisation but rather choose for a combination of multiple visuals to highlight different parts or perspectives of the data, often combined in a dashboard"[117]. For the dashboard, Shapley values serve as the main source, then presenting a combination of visualisations and text. The dashboard, by data scientists, is described as follows: "We have developed a dashboard and we've gone through a design process first our team and then a few domain experts. And then out of it came a dashboard. The domain experts seemed to like the dashboard. But it has some features that they thought are not useful at all and we, as data scientists thought, well, you gotta have this right? And they were like, I'm never going to use it. Why should I use it? But there were other things that they found useful" (B.2.1).

Another dashboard mentioned, is a dashboard for data scientists with the aim to increase the use of Explainable AI methods in their model development process. For this dashboard, the developers indicated that they might check feature importance for instance, especially if the models used are complex, but they also indicated that it is often easier or quicker for them to run the code directly, than to use the dashboard (B.2.3, B.2.6).

Debates on the dashboards

It is important to note here, that both dashboards are not structurally used. Reasons to explain why the explainability tool for domain experts has not been embraced, are: "The domain experts are actually not so keen on diving into all of it" (...) "We think, hey this is very helpful, and they think, this is completely not interesting" (B.1.1) or " And then all of a sudden Explainable AI came as a solution. But that does not necessarily make it useful for the domain expert, because it makes the model itself more understandable, but for the domain expert to understand the SHAP visualisations is another story (...) Within the pilot, it is the idea that they will use the dashboard, but it has been difficult" (B.2.2). However, another data scientist wonders if visualisation is really the problem, or if it is the ML techniques themselves. "It's such a complex technique, can you really simplify it more? Often it's the problem that you can make it very simple, to the extent that everyone can understand it. But it is the question then, if it is still represents what lies underneath? (...) So, perhaps you need to use easier techniques. Sometimes I wonder if we should not make the choice between explainability or accuracy of machine learning, and perhaps go back to decision trees, or rule based, or something else simpler (B.2.4). Next, reasons mentioned for the sporadic use of the dashboard for data scientists, are: "Well, I think everyone here programmes a bit in their own way. You need to adjust too much to use the dashboard. It is easier just to go to the source code of the dashboard, and then to apply it to your own model, instead of loading the data exactly the way the dashboards wants. That takes more time" and "I have a model and then you look at feature importance, a bit case-wise, right? Is it logical that the score is high or low? The SHAP values, additionally, are mainly to also adhere to the ethical standards, but I already had full confidence in the model. So the XAI metrics we use as such, have not contributed as much" (B.2.2; B.2.6).

Review framework

The review framework translates ethical values, as gathered by the public body to a set of guidelines. These are largely based on the values defined by the High Level Expert Group of the EU mentioned in section 2.3.1, including looking at Explainable AI. One of the respondents explains: "The current AI changes trigger resistance. And we now know that you can go off the rails as government, look for example at the 'toeslagenaffaire'. We want to prevent that of course. Therefore, we have been looking at the different review frameworks out there and like that, we came with a set of guidelines, also following the European Union as kind of a pre-selection for the AI Act to come" (B.2.2). The review process includes people from different angles, amongst which a data scientist who has not made the model, often someone from outside of the organisation, a domain expert and the idea is to potentially involve ethics experts as well.

Debates on the review framework

The review framework is mentioned a few times by the managers, and at least once by the majority of the data scientists. A domain expert mentioned 'sessions' in the context of exchange of ideas and reviewing. However, they are somewhat critical, saying to miss more practitioners in those sessions, and that the sessions are too technically oriented and less so from a domain expert's point of view. ”We have had a few sessions about AI and Machine learning. But if I look around me, I do not see anyone who has ever executed an inspection. So you can develop all these models and theories, but still you need to have someone from the ground present. And that is not solvable just by shadowing a domain expert just once” (B.3.3). Their reflection on the meetings, in which explainability is a key element, was ”We have all these sessions with post-its and stuff. It feels very much as the specific world of AI”.

The review framework is said to be helpful for managers, mainly for governmental questions of who decides over the algorithm, and which steps are necessary to adhere thereto? However still, they are only guidelines and still open to interpretation, evaluation and conversation. ”Of each of these questions we have asked ourselves, what do we know? What do we think? And to make sure every model reaches some sort of maturity level, we designed the review process (...) However, you also need to remain alert. If one person says, well this part is not interesting, you need to evaluate it, do you then also design your model differently?” (B.1.1).

A data scientists adds that it is still a balancing act to make the review process both technical, as well as understandable to non-technical people involved. ”About the review process, we have thought about it a lot, but I haven't quite figured it out. We have a conversation in the group and certain trade-offs are discussed. But say, if a certain choice is much focused on 'Which metric do we used to optimise?' That has a lot of impact on how the system works. Are you going to minimise your false positives for instance? However, that conversation is very technical. How do we get a domain expert or a manager to say, well I actually think it is more important to focus on the false negatives? Many of these technical pieces, if we write them down or discuss them, people think, whatever. However, how do we translate it, to include the managers and decisions domain experts in these choices? Since the impact of the choices is big” (B.2.6).

5.2. Key-actor’s (Information) Needs

Where previous section looked into the role of XAI, this section starts to explore the possibilities of XAI. Once the information needs of stakeholders are clear, one can tailor the explanations accordingly. One might notice that many of the information needs are in line with the challenges mentioned, though it is not a one-to-one mapping. Nevertheless, beyond information needs, it turned out that the respondents voiced other needs as well. Even though these might not be solved by explanations, the insights are still valuable for improved AI implementation within the public sector, and mentioned as such, starting with the caption *beyond information* in the summary and headings below.

5.2.1. Overview information needs

The respondents were asked which *information* would be desired to achieve an improved AI-system scenario. This translates into the information needs per respective actor-group. Since numerous information needs are mentioned, this section starts by giving an overview per actor-group. Note that some of the information needs could be clustered together. However, making the needs as concrete as possible will give more concrete insight into which explanations need to be given. Also, exhaustiveness is not realistic within the scope of one research, therefore the information needs serve as a starting point, rather than a definite framework.

Actor	**Information Needs**
Manager	1. Information to estimate the impact of a model 2. Information about bias and profiling 3. Information about the data heritage, availability and use 4. Information in the form of examples. What are the opportunities AI can offer? 5. Information to have a mutually understandable conversation 6. Information to foster trust 7. Information to increase awareness and an urgency to be explainable 8. Information about AI generally *9. Beyond information: build capacity and improve educational programmes* *10. Beyond information: agreements*
Data scientist	1. Information on which explanations to give to whom 2. Information on how to present the explanations 3. Information on (ethical) responsibilities 4. Information to understand the users *5. Beyond information: access to domain experts & conversation* *6. Beyond information: sufficient (building and test) data* *7. Beyond information: ownership and usefulness of the work* *8. Beyond information: freedom to be creative* *9. Beyond information: prevent to get lost in a paperwork jungle*
Domain expert	1. Information/’Proof’ that a model works 2. Information to make well-informed decisions 3. Information and recognisable examples to know the opportunities of AI 4. Information on judicial issues 5. Information about inputs and outputs of the AI model 6. Information (provision) about the decision-context *7. Beyond information: timing of the model outputs and explanations* *8. Beyond information: the feeling that the model contributes* *9. Beyond information: dedicated time to learning* *10. Beyond information: integration of different information sources* *11. Minimising information: preventing information overload*

Table 5.2: Overview information needs, mentioned per stakeholder.

5.2.2. Managers

The managers mentioned various information needs; ranging from information around the model, e.g. bias considerations, to information around the process e.g. achieving trust or mutually understandable conversations. From the interviews with managers, ten (information) needs have been discerned.

1. Information to estimate the impact of a model

Generally the managers (4/5) thought it was not very important to know the exact workings of the model, but it was rather the process or the impact they are concerned with. Given that some of the tasks of the studied cases have societal relevant- but also potentially harmful outcomes, managers expressed the importance of gaining insight into the impact of the model, though the managers did not always know where to start assessing such impact. It was voiced that even after impact assessment, the use of a model's outcomes may have unexpected or unwanted consequences still, but understanding the impact is considered a way to contain the negative impacts. "The details of a model are not necessary, but what is important, is knowing what the impact of a model is. Once you know if a variable is important or not, that is nice, but you still do not know what it means" (B.1.1). One of the management says: "When I speak at a management level, I often think they do not understand the bigger picture, and thus also not the consequences" (A.1.2). Additionally, a respondent said that estimating the model's impact "requires an interesting game on the content; to be informed about the newest advancements and that you can keep up" (B.1.2).

2. Information about bias and profiling

Already, whilst discussing the earlier mentioned perspectives on XAI of managers, information needs seeped through. Respondent A.1.2 said they wanted more transparency with regards to understanding "the black-box". Both information about the assumptions of people, as well as unwanted, potentially discriminatory, bias of the model was desired. "I want to know that we are not profiling on unwanted characteristics, say the size of one's shoes or the colour of their eyes, that the model starts to zoom in and finds more and more things based on these characteristics" (A.1.1).

3. Information about the data heritage, availability and use

The manages want insight into the data for bias and profiling purposes, moreover they want to know if the data is available at all - "Is the data available? That is one of the preconditions" (A.1.2), and transparency about which data is used. "In the context of explainability, it is important to know: which data do you use?", "I want to see what goes into the model, a list so to say. So which of the data did you use, and which information did you ignore?" ((B.1.1; A.1.1).

4. Information in the form of examples. What are the opportunities AI can offer?

The managers like to see examples or best practices of the use of AI, to know what an AI-model could contribute to the organisation, or to make the choice to initiate such a model. Also, to further explain the workings of the algorithms to the domain experts, including in earlier stages of development, they say examples help. "You need to explain the main points of what a model can do, so you can do that in text or it is often nice to have an infographic and show people visually how it could work (...) If you get started with something new, at some point you also want to show certain success of course. If you can show small successes, you can convince people of the benefit (B.1.3).

5. Information to have a mutually understandable conversation

What should we do, if we want to move towards including the understandings, references, values and norms of people into the model and into the use of the broader AI-system? Those are questions mentioned by all managers. "The data scientists can have the conversation amongst themselves, but the domain experts cannot have that explainability conversation yet", is another remark made about the need for conversations (B.1.1). "Very quickly, is becomes just every person for themselves" (B.1.2). "So we need to converse with people and find out how this could work for them" (A.1.2). Finding out then, not only to get people together and talk to each other, but also knowing how the conversations can be held in order for all parties to consider them both beneficial and understandable, was a much-mentioned managerial gap in information.

6. Information to foster trust

The desire for meaningful conversations was often accompanied by the wish to know how to foster trust. "I said to my management, yes well, (...) what it boils down to, we can make nice instruments and automate things (...) But basically what we need is to know is, how are we going to build that trust? How do you work together with people, but certainly also with the system?" (A.1.2). Another managers talks about wanting more information about the deeper layer behind trust. "I think trust is an emotion, and that we need to address it as such (...) Just telling a domain expert 'don't worry everything will be fine', does not cut it" (B.1.2).

7. Information to increase awareness and an urgency to be explainable

A manager wants more information to convey the urgency of Explainable AI, both for data scientists and for domain experts. "If I am missing any information? Yes, daily. In the context of explainability, there are a lot of technical options, you know. But how do you make sure that the data scientists have the feeling; I need to do something with it? And how do you make sure it becomes common shared knowledge? For the domain expert I find that even more difficult. We find that there should be explainability, but they are not always open for it" (B.1.1). They add that explaining in itself is a sensitive topic: if you are explaining the algorithmic system, how do you make sure you do not tell the domain expert how to do their job?

8. Information about AI generally

Previous chapter mentioned that people in the organisation are not always sufficiently knowledgeable about the topic of AI. Three managers came back to the point whilst talking about information needs. The managers said they themselves have an idea or understanding of the topic of AI, because they are interested in it and a few of them are the dedicated manager of an innovation team. However, they see an information need where it comes to other managers and the wider organisation. "See, if you look at creating understanding, you do not only need to look at explainability, but also look at understanding about what AI is in general. I can create the conditions, but our environment also needs to understand what it is" (A.1.2).

9. Beyond information: build capacity and improve educational programmes

Beyond information, the managers mentioned several needs for improved AI implementation within the context of explainability. The first need beyond information, is that of capacity building and improved educational programmes around AI-systems. "Currently, if you look at the general on-boarding process, data-driven working is not included. Now I know data-driven is not equivalent to AI, but even data-driven is not included. I find that a pity. My dream would be that from now onward, affinity with data is part of the vacancy procedures" (B.1.1). The need was raised to capacitate the domain experts to use AI. "They do not need to convert into a data scientist, because they are domain experts, who are good at other things. But you need to have a conversation about capacity building; how do I make a decision if I use AI? And how do I deal with mistakes I make in that process?" (B.1.1).

10. Beyond information: agreements about responsibilities, ethical considerations and maintenance

The managers talk about the need to come to agreements, on the topics of responsibility and accountability, ethical considerations and maintenance. As mentioned in the challenge of the responsibility gap in previous chapter 4.2.2, a general idea of responsibilities is recognised amongst the actors, yet there is no formalisation of the topic, and some of the responsibilities are not agreed upon by all actors. One manager mentioned that the fear of being held responsible might withhold some domain experts to use the model's outputs. Others said it is a topic that will settle with time. Either way, within the context or responsibilities, the managers indicated a need for clarity on responsibilities, and preferably coordination and agreements on the topic.

Next, a need for agreements on ethical considerations is mentioned. Where impact broadly, and bias and profiling specifically were already mentioned, so was the need to establish agreements on other ethical implications of the model, such as transparency and privacy. For instance, you do not want to spread information that might enable decision subjects to change the outcomes of the model, at the same time you want to be transparent. "If we want to be as open as possible, in fact it is even our task as government to be transparent, it is a trade-off, between openness and our legislative task, that is often made implicitly. It would be good to make that explicit" (B.1.1). The need for critical reflection and review is mentioned, on top of the agreements: "Asking each other critical questions, an open culture where you can question each other, also including the domain experts and people from outside. In such a process I experience the largest safeguarding" (B.1.2).

Third, in line with the challenge that maintenance is currently not assigned impeding the go-live of models4.2.4 , managers mentioned the necessity for agreements upon maintenance, but also for carrying the potential changes to the model, or ensuring structural evaluation.

5.2.3. Data scientists

1. Information on which explanations to give to whom

The data scientists described the need for clearer and/or more information on which explanations are expected or helpful, both from managers as well as domain experts. From management, there is a wish and even outreach to ask management which explanations they would like to have. "We asked management, well what do you need from us? Which information or explanations can we give you? But I don't know if they know yet, how much information they need" (B.2.1). For the domain experts, the question remains; how does the domain expert get the information they need to do their job properly, without overloading them with information? One of the data scientists describes: "You have to simplify things, but you have to make sure the information you show is succinct as well. You can't show too much information" (B.2.3). The mere reason to dedicate several MSc researches to it, indicates a desire to improve information on the topic. Even though the information level is increasing, still questions around the topic remain, such as: "How do you get people at the level where I think their responsibility lie, and they can make informed choices about it? (...) Even though we are aware, often the conversations remain technical, so how do you translate these?" (B.2.6).

2. Information on how to present the explanations

Next, once the data scientists know which explanations are desired, how do they best communicate them? By know, it is known that a scatter plot or other plots are not the best way to do so. But what is? "We already wrote an instruction manual, we wrote a bigger manual, we organised a session and we created a dashboard. What else should we create?", "First of all, the question is not always clear. But next, the domain experts do often not know what options we can offer. They might simply ask for a top 10 in list-form, but there are many more options to present it more clearly, of course" (B.2.2; B.2.4). Another data scientist adds that even though such information is desired, it might not really be the job of a data scientist to do so. "Sometimes, it is just about data visualisation. And that is difficult. We have people partially working on it, but it is not really our job. We don't really have an interaction designer or UX designer in our team, and I don't really know if that is the correct place, but it would be good for the organisation to have that knowledge" (B.2.6).

3. Information on (ethical) responsibilities

Linking to the latter point, when is explainability the responsibility of the data scientist, and when is it the responsibility of others inside the organisations? More clarity on that would help the data scientists. Most of the data scientists feel a strong moral responsibility, often with the rationale that otherwise they would not be working for a public sector organisation. They regard it as their task to understand the model themselves and be able to explain it to others if asked for. Often, they see it as part of their job to provide explanations of the model to future users, though that slightly differs per data scientist. But if they are also expected to do so, and what the boundaries are of being in charge for the explanations, or for other ethical considerations such as fairness, is not clearly defined. "The question is, who should do it? We don't really have an answer to that. Right now, our team has the most expertise to deal with ethical considerations, and therefore we do it, because it has to be done. Even though no one really likes to do it, it is important" (B.2.4). Multiple data scientists mentioned some form of the following point: that there is only so much one can do with their time, and even though all of these tasks are considered to be important, they cannot all be done by one person. Therefore, information on which are the ethical responsibilities expected as data scientists, would help improve clarity and focus.

4. Information to understand the model users & conversation

Every data scientist mentioned that more information about the domain experts' way of working and thinking would be beneficial. The data scientists currently in the pilot phase, stress it as one of their main information needs. "I would like to know what happens in the heads of domain experts. And why they do or do not cooperate for instance" (B.2.2). The need was raised to know how to make explainability not just about explaining the model, but also part of the mental world of the future user. To this end, the data scientists need insight into the current mental world of the future user. Moreover, a data scientists expressed the wish to know when the information is really understood. ”How can we check if the information that we thought off, that we want to convey, is also relevant to them? And that they really understand it?” (B.2.6). One way to address this, often mentioned by the data scientists, is through conversations. ”Sit down and have a few good meetings on what is going on. Hopefully, having a third party there would be amazing, someone who knows a bit about both worlds, so not really a super data scientist, not really a super inspector. Something like an inspectascientist” (B.2.3). Another way to gain the information, is by shadowing the domain experts in their work, perhaps whilst doing the pilot to see how they work with the AI-models and where there are gaps in knowledge on either side. Three data scientists indicate they have shadowed in the developing phase of the model, but never during the pilot phase.

5. Information for sufficient (building and testing) data

A few issues around the data were mentioned, and the desire to improve them. ”The data quality in and of itself is an issue. Real data is always a mess, that's just how it is. We are really transitioning to a pen and paper way of working, to a automated and information-based, data-driven way of working. However, there is not enough data. There is no big database that is structured. Though it should exist (...) We need to first solve getting our hands on the data and try to figure out what we can or cannot do with it” (B.2.3). Moreover, ideally the need to have randomly sampled data was voiced (C.2; C.3). Currently, the models are based on data where decisions have already implicitly or explicitly been made. However, though random sampling is desirable, respondents countered that domain experts are short on time as is, let alone if they need to add randomly made decisions to their tasks with the purpose of generating input to train and test AI models.

6. Beyond information: access to domain experts

Closely connected to the understanding of users, is the access to the time of domain experts and the physical access to either shadowing or having a meeting with them. ”I would need that more time is allocated, more time to brainstorm with each other, to talk and think with us about the models we make. Now we really need to hunt the team and the people who can make time for us. It is really just less than a hour per week we need” (B.2.5). The data scientist explicitly added that explainability is only a small part of getting the domain experts involved with AI-systems. ”They are quite far removed from us. Oftentimes, we build a solution and we bring it to them, they do not come and get it. (...) We need to convince them that the model can help them. (...) But it is just not their world” (B.2.5).

7. Beyond information: ownership and/or the idea that something will be done with the work

The data scientists indicate that they would like to see more ownership of the models, or to get the feeling that their work is getting used at some point. Having a problem that is well-defined or having someone take ownership of an organisational problem then to be addressed with the help of AI, was mentioned by four data scientists as a key-factor to success. ”The projects that go well, are if someone on the other side clearly wants it. But we don't always have that. Partially because they are just too busy, or maybe because they are not ready for it, or because we have not been able to show that the models are of added value, or because they have more important things to do” (B.2.6). Another data scientist adds: ”Sometimes I have the feeling the domain experts have quite a lot of power. If they do not feel like it, then people simply listen to it (...) Sometimes you hope that a director or a team lead takes the lead to make sure the change will really happen” (B.2.2).

8. Beyond information: freedom to be creative

The data scientists point to the fact that, where on the one hand it would be good for the implementation of algorithms to have a clearly defined problem and ownership, on the other such choices might also mean less creativity and freedom. The existence of an innovation team, is to be creative, to have ideas and think of solutions that might not have been thought of yet. Where some data scientists preferred a clearer-defined direction with the knowledge that there is a higher chance of implementation of the algorithms, others expressed a desire to remain in charge of their own work.

9. Beyond information: prevent to get lost in a paperwork jungle

The need for ethical checks is acknowledged, that much we know. However, a few times the word 'bureaucracy' was negatively used in this context. For instance, a fear is expressed that the the introduction of the AI Act will result in a pile of paperwork, in a jungle of check lists and registers. One data scientist even thinks it might work counter-productive, in the sense that it discourages people to be honest and critical about their work. "I am not a big fan of the upcoming AI Act, because it has a heavy documentation load. For better systems, there should be freedom to remain critical. Now, anytime you would write something critical, if you know five others need to approve it who might each stop your work, of course you will not write down vulnerable and critical things about your project" (B.2.6). The term 'paper tiger' was mentioned here. "I am not a big fan of the AI Act that is coming. I think there is a large documentation duty there. For instance, if I recall the procedures of the 'AVG' (stands for 'Algemene Verordening Gegevensbescherming' which is the Dutch privacy law). In practice it means filling out a ton of forms and lists without any real change to the privacy impact of the system. The only thing we do is writing things down, but the impact on the citizens in the end remains the same (...) I really believe in these ethical checks, but I think you need to put it in conversations between people and not in a checklist" (B.2.6).

5.2.4. Domain experts

1. Information/'proof' that a model works

Domain experts would like to see results of the model, and get an idea if they can trust its outputs. "So you need to show, this works and this doesn't. Right now, we do not have any material for comparison, which makes it difficult. If you would have the scenarios next to each other, one with and one without the AI-system, then you can compare. If the model works better, it works better. If it doesn't work better, you'll see it naturally" (B.3.2).

2. Information to make better-informed decisions

Every domain expert expresses a desire to make the decisions as well-informed as possible. For some, that means moving towards a more preventative strategy in identifying risks. "The motivation for me to work with the models, is the current time pressure and the desire to know risks at the forefront. The idea then, is to move away from an ad-hoc situation and towards a situation of knowing and informing decision subjects" (B.3.3). For others, it means widening their perspective, to prevent a tunnel vision in their decision-making. The domain decision experts expressed the need then, for the information gained from the AI-system to contribute to their decisions and widen their angle, therewith making a better-informed decision. A side note is placed here, that the information obtained by the AI system should become something they need to follow, but rather a tool or help in their own process.

3. Information and recognisable examples to understand opportunities of AI

Next, multiple domain experts mentioned that explanations in the form of examples would help to understand both how the model works, but also to themselves identify opportunities where AI could be of help. "To give an oversimplified example; if the model never finds anything in a blue container, than you can skip all blue containers. Those findings will bring you to ideas. Or maybe the outcome will be the model needs to be adjusted, that is also possible" (B.3.2).

4. Information on how to deal with outcomes outside of their current jurisdiction

Sometimes, the model may present outcomes the domain expert cannot follow up upon. One domain expert who has worked within a pilot, indicated they would like information on how to deal with such situations. "If you go to a decision subject indicating that you have used a model, sometimes you get questions asked by lawyers. Also, sometimes we get information but we cannot do anything with it because it is outside of our jurisdiction (...) Also, if you go to a decision subject, you want to indicate why you are there". The domain expert indicates that legislative expertise would help in such instances.

5. Information in the form of an easy overview of inputs and outputs of the AI model

Providing an overview of the inputs and outputs of the model, is regarded helpful by multiple domain experts (B.3.2; B.3.3; B.3.4). Knowing which data is used and where it came from, and how it was translated is considered important, to gain insight into the model (B.3.4; C.2). "It is also a societal question. Of course I want all the information to come to me, but we need to be able to estimate how far we want to go with that".

6. Information (provision) respecting the decision-context

The domain experts indicated that, not only do they need information, they would also like to provide information. ″I do not have a good idea about AI, but at the same time I feel that the people from the innovation do not always have a good idea of what I do. Well, I think there is only one solution then, that is to walk along and observe each other″ (B.3.4). ″It will also help with transmitting the message. Sometimes just using different words, different nuances can help a lot″ (B.3.3).

7. Beyond information: timing of the model outputs and explanations

Not all the decisions need to be made under time pressure, but some certainly do. Moreover, agreements with other organisations have been made concerning timing. Therefore, the domain experts say they can run a model for a while, but the information needs to be provided relatively quickly. You cannot say ″hey, well in three days we can come back to you with our explanations″ (B.3.2).

8. Beyond information: the feeling that a model contributes, instead of taking valuable time

Right now, the domain experts do not always see the added value of an AI model. One of the domain experts goes into this point more in-depth: ″The question is also if the domain experts are always ready for it, it takes time and effort. That means the domain expert should not be overloaded with models that have not been developed yet, or will not be implemented due to a lack of knowledge, data or money. Often, from our side, we indicate things, but then we don't see them back in practice. It is difficult then, to keep the domain experts enthusiastic and, with the limited time and pressure, also from outside of the organisation, it is important to ensure the domain experts experience as little burden as possible of things that are not necessary″ (B.3.3). Affiliated with the feeling that a model is of added value, is keeping room to use the model flexibly contingent on the situation. ″There needs to be room to think for yourself, that is really really important″ (B.3.4).

9. Beyond information: dedicated time to learning and model-inclusion

Throughout the results chapters, the need to have dedicated time to learn about AI and to be involved in the model implementation process, seeped through. The challenges of lack of priority, understanding and inclusion of the domain experts, limited resources, were all accompanied by a request for more time. Some respondents argued that the time invested will pay back naturally, since decisions are made more efficiently. Nevertheless, domain experts and a few other respondents argued that even if this might theoretically be the case, reality shows that the process towards reaping such potential time-profit, is long, time- and energy-consuming. Without specifically dedicated time and tasks, or the allocation of financial resources towards the development and adoption of AI then, the decision-makers will not magically dedicate their limited resources towards a technology they are not intrinsically familiar with and/or curious about.

10. Beyond information: integration of different information sources & usability

Whether in the form of a dashboard or whether communicated differently, the domain experts express the need to have a more integrated presentation of different information sources. It would increase the usability of the AI system. Three quotes expressed by the same respondent in this regard, were: ″My need, in terms of information, is actually less manual searching (...) Preferably, I would like all my information on a platter″, ″it would be nice to have a assembly point for all the information″, and ″I am fan of dashboard functions, if they are integrated in the main programme you work with″ (B.3.4). Right now, the domain experts say they have multiple information systems, both on- and offline to deal with. Adding yet another one is doable, and ″a matter of time to get used to it″, however it adds complication and is not preferable (B.3.2; B.3.4).

A similar point is made for the information provided by people. ″The data now, is in the dashboard, but it doesn't help that the innovation lab works with many students, and then implementation lies in a different team, with different knowledge and tools″. Moreover, a domain expert says to need to deal with many sources of information on the ground as well, who are not always willing to provide information or well-aligned with the goals of the studied organisation (B.3.1). Having more channeled and better-communicated information sources would help.

11. Minimising information: beware of information overload

The domain experts need succinct information and comprehensible information; they warn against an information overload. The initial reaction of a domain expert was the following: "Which information I would need? Please do not add anymore" (3.4). The message to beware of information overload was shared by all domain experts and a few data scientists ((B.3.1-B.3.4; B.2.1; B.2.2; B.2.5; B.2.6). "Too much information is really a thing, there's a balance between how much information you give them. (...) In our case we don't have models that decide on their own, there's always a human in the loop. I think also from a legal and ethical perspective this is called *menselijke waardevolle tussenkomst* ('valuable human intervention'). Because we work with high impact issues, it is not about telling them how the model works, but just giving them sufficient information about the model output that they can decide well, I am able to use this information, or rather not" (B.2.1). Here again, *how* explanations are given is important.

5.3. Chapter Summary

From this chapter it is important to take away at least three things: XAI has multiple purposes, XAI is targeted to different 'audiences' or actor-groups, and every actor-group has information needs, largely informing how the purposes of XAI could be further operationalised. Answering the second sub-research question, "What is the current role of XAI in the public sector?", the role of XAI is unpacked into three aspects: the perspectives on XAI, the purposes for XAI as mentioned by the respondents and the current use of XAI. The current use of XAI is limited in one of the studied cases. In the other case, it is addressed through the use of feature importance methods (SHAP), dashboards and a review framework. The current use of XAI (in the case which uses XAI) enables for better understanding of the models for mainly the data scientists, increased awareness of the topic, and the ability to improve the review process of models developed. However, there is still a gap in communicating the insights (e.g. of the Shapley values) to less technical audiences.

The respondents saw the importance and potential contribution of XAI beyond its current use. They mentioned several purposes for XAI; every actor starting with purposes for domain experts. Especially in one of the two cases, XAI was mainly mentioned as a form to improve understanding and usability by their potential users, but after further prompting of the researcher, the respondents explored XAI for their own respective actor-group.

Answering the third sub-research question, "What are the information needs, to improve understanding of the AI system, as defined per stakeholder?" the information needs for every stakeholder are researched (overview in table 5.2). For managers, the information needs are closely related to increased accountability on the one hand, such as estimating the impact of the model, information about bias and information about the data, and on the other hand to enable improved conversation and awareness. Explainability that would flow from these needs, is more about the apprehension of the larger system and the rationale behind it, than about understanding the 'machine' or details of the model per se. For data scientists, the first concern is to get a model working and, once they have, their question is how to know which explanations to give to whom. The information they require then, is often related to communicating and distributing their work to others in the organisation. For the domain experts, gaining confidence in whether 'the model works', knowing why a model needs to be used, understanding the possibilities AI can offer, getting guidance in AI-related judicial issues, and knowing the model's inputs and outputs are their main information needs.

However, the respondents clearly indicated that they do not only need information. They have other needs as well, *beyond information*, to improve the AI implementation process. Examples are capacity building, drafting of agreements and more allocated time and resources. Moreover, beware that there exists something like 'too much information' and 'too many rules'. Respondents warned for an information overload and the danger of getting lost in bureaucratic burden. Providing sufficient information people can interpret and use the models for their respective tasks, yet without overwhelming them, was considered key.

6
How could XAI contribute to AI Challenges

To connect the AI implementation challenges (chapter 4) as well as the XAI findings (chapter 5), this chapter looks into the last sub-research question: ”How could XAI contribute to public sector AI implementation challenges?” To answer this question, a few components discussed throughout this thesis are used, mainly: the purposes of XAI as ideated by the respondents (table 5.1), and the mapped public sector AI implementation challenges (figure 4.1).

This chapter starts by reviewing the purposes of XAI practice against those found in literature, and then views to what extent those operationalised purposes could contribute to AI implementation challenges.[1] The aim of this chapter is to provide a starting point for realistic conceptual mapping; reflecting the opportunities and importance of XAI to address certain AI implementation challenges, but also reflecting that XAI is not a miracle solution to all the existing AI-related troubles.

6.1. Operationalising XAI Purposes: from Literature to Practice

This section serves as an intermediate step to answer the sub-research question, making the XAI purposes from literature more concrete by comparing them with the interpretations of respondents (section 5.1.2). Going back to the literature, five purposes for XAI were distinguished: Model-improvement, Impact assessment, Control and accountability, Decision-improvement, Contestability, and Learning and Evaluation. For every ’audience’ or type of actor, a main purpose of XAI was distinguished from literature; the manager benefiting from XAI for impact assessment, control and accountability, the data scientist from XAI for model-improvement, and the domain expert from XAI for decision-improvement. All of the actors were identified to benefit from learning and evaluation. Already, it was acknowledged that contestability is important, though not the focal point of this thesis, since the decision subjects are not studied. Therefore, for the remainder of this chapter contestability is taken out of the analysis.

If we look at the purposes as distinguished from literature, and link them to the purposes to use XAI as defined by the interviewees themselves, we can see that overarching XAI purposes are largely similar to those distinguished from literature (see table 6.1 below). Nonetheless, they are not strictly separated between the audiences. For instance, ’clearer understanding of the responsibilities’ identified by the data scientists, might contribute to model-improvement, yet is more logical to contribute to accountability. Some of the other XAI purposes from practice could fit into more than one category, such as ”communicate to the decision subjects and public’ (both accountability as well as contestability), ’know the model’s bias’ and the biased assumption from people (accountability as well as model-improvement as well as decision-improvement), ’gaining insight into the data used’ (accountability as well as model-improvement as well as learning and evaluation).

[1] Where both previous chapters closely followed the interviews, this chapter uses more interpretation and insights from literature.

Consequently, though coupling the purposes of XAI to one actor group helps to keep oversight, when breaking them down for the respective actor-groups, they will need to incorporate more than just 'impact, accountability and control' for the manager, 'model-improvement' for the data scientist' and 'decision-improvement' for the domain expert.

XAI Purposes from literature	**Actor**	**XAI Purposes from practice**
Impact, Accountability & Control	Manager	- To be accountable - Know the model's bias - Know the biased assumptions of people - Communicate to decision subjects and the public - Estimate or act upon potential mistakes made - Explain the model towards other managers - Gaining insight into the data used
Model-improvement	Data scientist	- Insight into more complex models - Fine-tune the model - Verify model's assumptions - Understand (technical) unwanted biases of the model - Clearer understanding of responsibilities
Decision-improvement	Domain expert	- Enable better-informed decisions - Understand assumptions and criteria of model - Understand model's input - Demonstrate that the model can be trusted
Learning and evaluation	All	- Creating a critical and open process - Verify process- and model assumptions - Take away current tension of the 'unknown'
	Manager	- Follow the right steps in the process
	Domain expert	- Show how to work with the model - Increase use of models - Recognise their ways of working in the model

Table 6.1: Coupling XAI purposes from literature, and their corresponding actors, to XAI purposes in practice.
The XAI purposes from literature can be found in section 2.3.3. The XAI purposes from practice in section 5.1.2 and table 5.1.

*Note that many of the XAI purposes identified in practice match the overarching XAI purpose from literature, but not all. Some purposes from practice would fit better in one of the other categories from literature, and some might serve multiple purposes at the same time.

6.2. Operationalising Opportunities: from XAI Purposes to AI Implementation Challenges

This section looks into the opportunities XAI can offer for the discerned AI implementation challenges (figure 4.1). Following the purposes distinguished earlier, though taking out learning and evaluation, since it was often mentioned throughout the conversations and less easily discernible as such, every XAI purpose is linked to the AI implementation challenges it could contribute.

The practical break-down of XAI purposes served as a first intermediate step. The second intermediate step, is the table below, connecting the different XAI purposes to specific AI challenges. The idea is that XAI purposes offer the opportunity to address some of the AI challenges at hand. Note here, that some of the challenges could be improved by multiple purposes of XAI. Also, note that some of the relations are more evident than others. For instance, one can debate if being accountable only contributes to a better estimation of the impact of the model and/or to tighten the responsibility gap, or also to fairness. Therefore, it is important to know that these claims are not meant to be exclusive, but rather to offer a framework to start conceptualising the connection between XAI purposes and their contribution to empirical AI implementation challenges.

XAI Purposes literature	**XAI Purposes practice**	**AI Implementation Challenges**
Impact, Accountability & Control	- Be accountable - Estimate/act upon potential mistakes - Estimate model's bias - Gaining insight into the data - Biased assumptions of people - Communicate to outside world - Explain towards other managers - Understanding responsibilities	- Uncertain model impact & -Responsibility gap - Uncertain model impact - (Un)fairness - Unclear 'why' for AI & - Data considerations - Ethical considerations in modeling - Ethical considerations in modeling - Lack of priority - Responsibility gap
Model-improvement	- Insight into complex models - Verify model's assumptions - Fine-tune the model - Biases of the model - Gaining insight into the data used	- Difficult modeling choices - Capturing 'soft' mechanisms in 'hard' models - Modeling choices & - Capturing mechanisms - Data: historical bias - Data: sources, insufficiency and bias
Decision-improvement	- Enable better-informed decisions - Know assumptions/criteria of model - Understand model's input - Know if the model can be trusted	- Usability & - Novel ways of working - Resistance to change - Ethical considerations - (Dis)trust

Table 6.2: Table to conceptualise how XAI purposes could contribute to AI implementation challenges.

*A disclaimer is in place: the relations are not always one-to-one and subject to interpretation. The aim however, is to provide a conceptual framework, and understand how XAI purposes could be realistically translated to important AI implementation challenges.

6.2.1. Conceptual framework: XAI addressing AI implementation challenges

The last step of this research, is to visualise the framework originated above. Here, besides knowing which AI implementation challenges can be addressed by XAI, the AI implementation challenges which cannot be address by XAI, or at least not directly, are depicted as well. Moreover, the challenges are shown along the lines of the AI Implementation life-cycle phases, with the reasoning that it can help public sector organisations to make XAI more practically applicable within the larger AI implementation process.

The intermediate steps above are taken out of the figure, though know that they clarify the relations drawn. Each of the purposes has their own colour, and so do the different types of challenges (for the types of AI challenges, the same colours are used as before). The disclaimer mentioned in previous table is relevant for this figure as well: the mapping is open for debate and interpretation. The visual framework serves as a conceptual model, to see which public sector AI Implementation challenges can be addressed by the discerned XAI purposes and which less easily so.

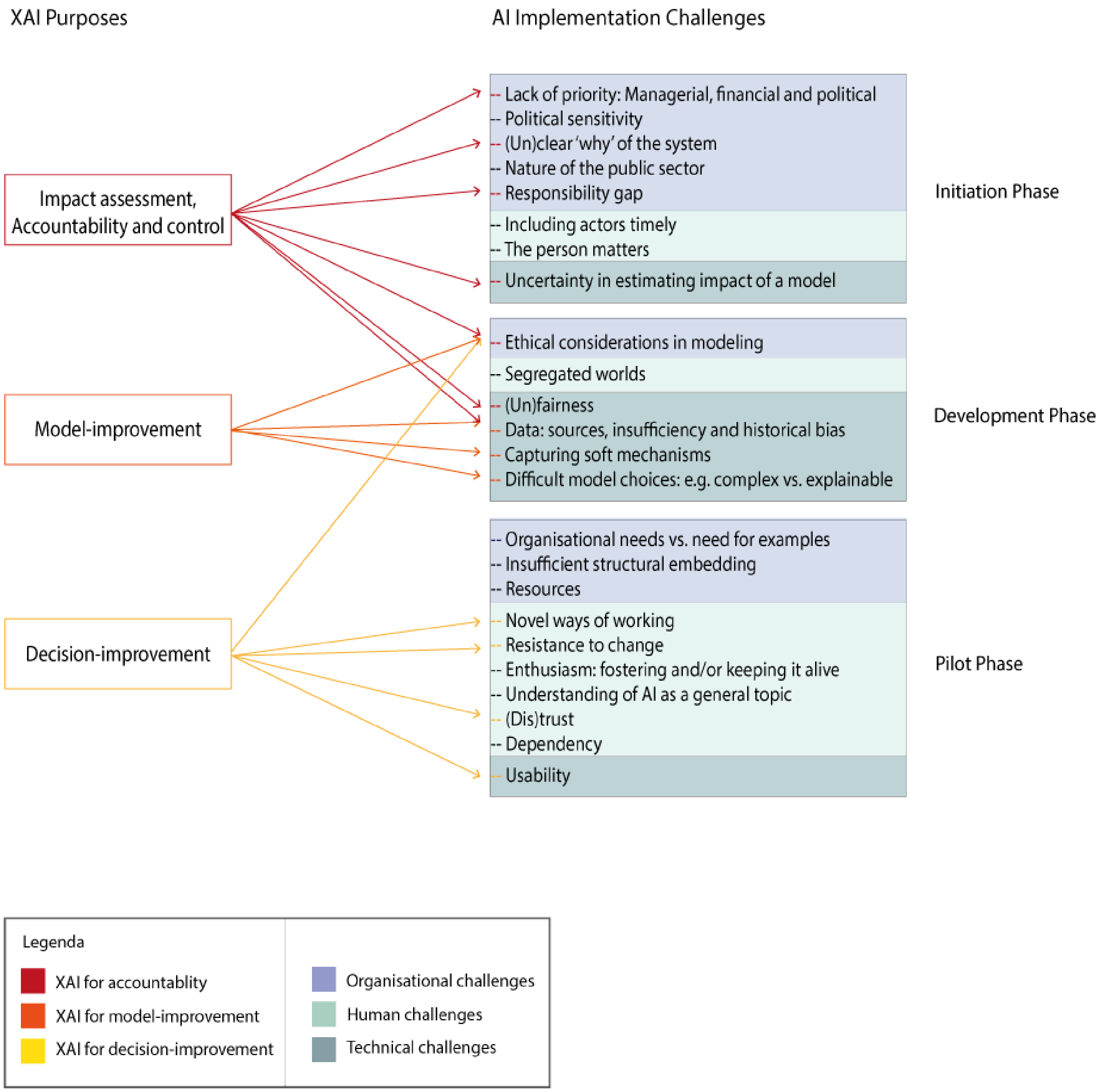
XAI Purposes
AI Implementation Challenges
Impact assessment, Accountability and control
Model-improvement
Decision-improvement
-- Lack of priority: Managerial, financial and political
-- Political sensitivity
-- (Un)clear 'why' of the system
-- Nature of the public sector
-- Responsibility gap
-- Including actors timely
-- The person matters
-- Uncertainty in estimating impact of a model
Initiation Phase
-- Ethical considerations in modeling
-- Segregated worlds
-- (Un)fairness
-- Data: sources, insufficiency and historical bias
-- Capturing soft mechanisms
-- Difficult model choices: e.g. complex vs. explainable
Development Phase
-- Organisational needs vs. need for examples
-- Insufficient structural embedding
-- Resources
-- Novel ways of working
-- Resistance to change
-- Enthusiasm: fostering and/or keeping it alive
-- Understanding of AI as a general topic
-- (Dis)trust
-- Dependency
-- Usability
Pilot Phase
Legenda
🟥 XAI for accountablity
🟧 XAI for model-improvement
🟨 XAI for decision-improvement
🟪 Organisational challenges
🟩 Human challenges
⬛ Technical challenges

Figure 6.1: Visual conceptual framework, showing: which XAI purposes can contribute to which AI implementation challenges, and which AI implementation challenges are less easily addressed by the XAI purposes.

*The mapping is based on the empirical findings within the public sector context of the research. Still, a disclaimer is appropriate: the relations between XAI and AI implementation challenges are open to interpretation and debate.

From the figure, reading from the XAI purposes on the left, at least three things can be noticed, and the same goes for reading from the right, from the AI implementation challenge:

1. Impact assessment, accountability and control is mostly targeted towards AI implementation challenges in the initiation- and development phases.

2. Model-improvement is mostly targeted towards AI implementation challenges in the development phase.

3. Decision-improvement is mostly targeted towards AI implementation challenges in the pilot phase.

4. All of the technical challenges as identified by respondents can all be addressed by one of the three purposes.

5. Some of the organisational challenges, especially in the initiation phase but also some in the development phase, can be improved with XAI.

6. Some human challenges in the pilot phase could be improved with the help of XAI, though not all and less so in the earlier phases of AI implementation.

6.2.2. Further analysis of the findings

Explanations with the purpose of impact, accountability, and control, predominantly address AI implementation challenges in the initiation- and development phases and predominantly address organisational challenges, though also some of the technical challenges mentioned. They might help to gain insight into the impact of the model, address a responsibility gap if the responsibilities are more clearly defined and debated, improve insight into the data which can both help for increased understanding in why we should or could use the model in the first place, but also to counteract historical bias encountered in the data. Also, they might help to explain the model to other managers and management levels. Right now, the respondents indicate that the larger the distance between (higher) management and the actual model development, the less involved they are with understanding the AI-systems. Creating some sort of mechanisms where explainable AI trickles up, so to say, towards higher management, might be beneficial to tackle certain challenges mentioned, including enlarging priority in the wider organisation. Nevertheless, note that there are other organisational challenges remaining, not addressed by one of the XAI purposes. Examples are; the nature of the public sector, structural embedding of the system, and the availability of (sufficient) resources. These are not dependent on understanding as such, but for instance on time, money, organisational structures, or politics.

Explanations with the purpose of improving the model, are most useful in the development phase, and for technical challenges. The XAI methods as researched traditionally, think of the surrogate models or feature importance mechanisms (e.g. SHAP) mentioned earlier, are considered helpful within this category. However still, the methods can be further fine-tuned and adapted to the specified purpose, and the exact challenge to be addressed. Also, it is important to keep thinking of these XAI methods for model-improvement, even if the models are considered to be 'relatively simple' or 'easy to understand'. Where a model, its details and assumptions might be easy to apprehend for its developer, it should remain so for other developers and relevant actors outside the development team as well.

Explanations with the purpose of decision-improvement, are regarded most useful in the pilot phase and for human factors. Note however, that for all the implementation challenges but for these even more so, *how* the explanations are given, is arguably as important as *which* explanations are given. Presenting the right amount- and level of explanations, and showing the model's input and assumptions, largely influence to what extent they help to tackle AI implementation challenges. The reason that this is considered more important here, than for the other purposes, is that human challenges first and foremost involve humans, therewith they are largely perceptible to people's minds and interpretations. Explaining the assumptions and criteria of a model are a good start then, but whether this should be in the form of text, visualisations or something else entirely else, remains unclear and requires further research.

Additionally, **one could use the information needs** identified in previous chapter, **to refine the XAI purposes**, and create types of XAI necessary for the respective actors. However, additional research is needed to see which types of XAI (e.g. distinguishing between rationale explanation, responsibility explanation, data explanation, fairness explanations, safety and performance explanation, impact explanation) fit best for the actors and information needs. Moreover, certain challenges need to be addressed beyond explanations of the model. For instance, by offering educational tools or on-boarding for new colleagues on general AI principles. And by acknowledging that people need to be included from the start, for instance through participation methods, or simply by working from the same location more often, to desegregate the different worlds.

6.3. Chapter Summary

This chapter identified the opportunities and limitations for XAI to contribute to various of the AI implementation challenges, answering the fourth and last sub-research question "How could XAI contribute to public sector AI implementation challenges?". The short answer is that XAI can contribute to some of the AI implementation challenges, but not to all, and that a conceptual framework might be of help, to realistically identify where XAI could be beneficial in addressing the AI implementation challenges.

To come to such a framework, intermediate steps taken were taken in this chapter: the XAI purposes from literature were compared to those in practice (table 6.1), the XAI purposes from practice were reviewed against potential implementation challenges they might address (table 6.2), and the XAI purposes were mapped to see which of the implementation challenges could be rather logically addressed by the purposes of XAI and which less so (figure 6.1).

Impact assessment, accountability and control were mentioned to be most important for managers, within the context of ethical considerations; for instance, understanding the biases underlying the model, knowing how to act if something does not go as foreseen or understand who has the responsibility for which part of the process. Model-improvement was mentioned to be most important for the data scientists, to improve- and fine-tune the technical workings of the model, but also to understand the model's assumptions and make sure unwanted biases were accounted for. Decision-improvement was mentioned to be most important for domain experts, to enable better-informed decisions, and understand the outputs of a model.

With the more elaborate understanding of the XAI purposes both from literature and practice, one can identify opportunities to address AI implementation challenges. Organisational AI implementation challenges within the initiation- and development phases, were mostly addressed by XAI for impact assessment, accountability and control. Technical AI implementation challenges in the development phase, were mostly addressed by XAI for model-improvement. Human AI implementation challenges within the pilot phase, were mostly addressed by XAI for decision-improvement. However, note the following points. 1). some of the AI implementation challenges could be improved by multiple XAI purposes, 2). the ways of shaping XAI will determine to what extent it really contributes to the AI implementation challenges, 3). the conceptualisation serves to enable further debate and discussion, and might thus be open to different interpretations, 4). there are public sector AI implementation challenges which are not easily addressed by XAI.

7
Discussion

Based on the literature, Explainable AI could be of help, or is arguably even necessary, to responsibly implement AI practices within the public realm [11, 94, 95]. Nevertheless, talking to respondents in practice, Explainable AI is not always on the forefront of people's minds. From the interviews soon, three things became clear: there is a gap between pilot- and deployment of AI implementation in the public sector (derived from chapter 4), XAI needs to be approached from a multi-purpose, multi-actor conceptualisation (derived from chapter 5), and XAI might be a solution to some of the current AI implementation challenges, but not to all (derived from chapter 6). Reflecting back upon these findings by setting them in the broader landscape of both AI implementation literature, as well as Explainable AI literature, is the aim of this chapter. Moreover, the chapter will discuss the limitations of the research, and point to directions for future research.

7.1. Positioning the Results vis-à-vis Literature

Where XAI is often presented as a solution for multiple AI implementation challenges, there is little empirical research into the effect of AI implementation in the public sector [11], including its opportunities and challenges. From the few studies conducted, the majority focus on the use of AI, assuming that AI has already been deployed [58]. The results from this research however, aiming at empirical understanding of the effects of AI implementation, revealed that AI implementation often lingers between the stages of a pilot- and deployment phase. Therefore, in need for public scrutiny of AI implementation, the research decided to take a step back from XAI, first reflecting upon the AI implementation challenges as currently faced in real-life public sector settings. To then see if the AI implementation challenges can be addressed by XAI. This section positions the three discussion points, just mentioned in the chapter's introduction, within a wider scientific debate.

7.1.1. Bridging the AI life-cycle pilot- and deployment phases

Even though many organisations are increasingly technically adept, the AI implementation processes for public sector decision-support systems tend to get stuck in a pilot or testing phase, rather than reaching full implementation and adoption in the wider organisation. Disentangling why so, various AI implementation challenges were identified (figure 4.1). Going back to literature, two important observations can be made: 1). many of the challenges less easily addressed by XAI (figure 6.1) are arguably related to AI adoption, and 2). a shared inter-actor language surrounding AI, or so-called shared 'mental-model', might be of help.

To start with the former, the literature review stated that AI adoption adds a layer to implementation, to ensure that an AI system is really embedded within the larger organisation [61]. The much-heard desire for mutual understanding and improved cooperation, are - predominantly human - factors that closely relate to adoption mechanisms. Many of these factors manifest themselves later on in the AI implementation life-cycle, often during the pilot phase. Note however that a life-cycle implies a circular process, providing both the opportunity as well as the necessity to iterate on the phases and their challenges. Arguably, if the human factors would be anticipated from an earlier phase in the process, including adoption mechanisms dedicated to knowledge creation, skills, and trust, they might be less of a challenge later on. Take the challenge of incorporating novel ways of working introduced by the algorithm; if the algorithm would be co-designed with domain experts, they might be more familiar with the novel way of working already, creating less of a challenge or resistance later on. Similarly, if general knowledge of AI would be included in the on-boarding or schooling trajectory of the organisations, the possibilities of- or reasons for AI are more clear once they reach the pilot phase.

Literature talks about the concept of 'mental models', both to improve the quality of explanations, as well as potentially lead to AI adoption. A mental model is a person's internal representation of the people, objects and environments they interact with [118]. Incorporating these social constructs of explanations, and creating shared mental models, could help as a way to foster mutual understanding, especially amongst actors with diverging views [12, 118, 119]. Since human challenges involve people - thinking and acting from their perspectives, rationale and feelings - a good start is to involve these different people to share their ideas from an early phase onward. It is not only the less technologically 'advanced', aware or inclined people who need to learn to trust the algorithms, but also vice-versa: it is listening to the expertise and practical implications offered by the different actors around the table, which need to be incorporated into the algorithms, for the AI-system to be worthy of trust.

7.1.2. Towards a multi-purpose and multi-actor framework of XAI

The literature review already highlighted the importance of approaching XAI beyond a model-based or technical approach. The empirical results confirmed this. More specifically, the need to acknowledge that XAI can serve multiple purposes, and making these more explicit, helps identify which types of explanations are necessary and for whom. Literature mentioned different 'audiences' in this light, each of which need different things from Explainable AI, with different purposes, but also in different ways [95]. These different purposes identified in the literature review were largely adequate to categorise the empirically found purposes for XAI. However, reflecting on the findings, the purposes of 'impact assessment, accountability and control' and 'decision-improvement' can be better operationalised, the purpose of 'contestability' is hardly thought off before deployment, and the purpose of 'learning and evaluation' is still rather blurry and not very well delineated.

In the literature review, each purpose was coupled with an actor who would predominantly benefit from explanations for that purpose 2.3.3. In practice, this distinction is helpful, though the lines are not clear-cut. Moreover, the need for multiple purposes is acknowledged in practice, yet the XAI methods used, such as surrogate models or feature importance mechanisms, are still largely technical, and mostly catered to data scientists and/or other technical audiences. Having an Explainable AI discourse especially targeted towards a). the domain expert, but also b). the manager, seems vital.

An Explainable AI paradigm for managers, is missing both from practice as well as literature. One might argue that review frameworks or legislation such as the AI Act are tailored for the managers, however, these remain procedural and offer little information about the choices, trade-offs, limitations and pitfalls of the AI-systems. Having the right level of detail of explanations for the managers to make such important choices, from the initiation onward instead of as an assessment tool post-development, enables better-informed and essential ethical, political and organisational debates. In turn, increasing impact assessment, accountability and control for the managers, with active responsibility for the AI system on the forefront, and improved public accountability along the way.

Note here that different stakeholders have different priorities, thus future research into negotiating certain trade-offs and making them explicit would be beneficial. In the literature review, a few controversies and trade-offs of XAI were mentioned [11, 12, 14, 94, 95, 111, 112]. Empirically, the trade-offs between an explainable model versus a better performing one, overconfidence versus calibrated trust and transparency versus privacy were mentioned as well, even more since privacy in the public sector is also related to politically sensitive and legislatively bound issues. Additionally, two trade-offs in practice showcase a chicken and egg dilemma; 1). people identify a need for examples from practice to start an AI-system, yet they want the examples to come from practice (starting from innovation vs. starting from examples), and 2). one needs sufficient priority and resources for explainability to have impact, yet the priority and resources will only be freed after impact is shown.

7.1.3. Addressing responsible public sector AI implementation beyond XAI

Beware that one does not stretch the term 'Explainable AI' infinitely. It is often a pitfall that, if one talks about multi-disciplinary, multi-purpose or multi-actor perspectives, they can encompass anything and everything. Stretching the concept too much, revises to being contra-productive in terms of applicability. This is a phenomenon described in policy science for instance [120], but it can be noticed in the studied cases as well; where an abundance of definitions and possibilities makes it difficult to translate the concept to understandable tools for all audiences. A first step then, is to call for specific purposes for specific actors, as stressed above. However, it also calls for acknowledging realistic benefits as well as limits of the concept, and the need for additional mechanisms beyond Explainable AI.

7.2. Limitations

The methodological limitations of this research mentioned in the methodology chapter, are important to position the findings (section 3.4). Two case studies are not enough to generalise the findings of this research, thus verification with other similar case studies, and perhaps with other types of public bodies will provide the ability to broaden the insights. Next, ideally one would speak to more domain experts, given their importance yet difficulty to reach, and to decision subjects who were not part of the research. Also, categorising people in a group itself, reduces the ability to incorporate their individual tasks, hierarchy and perspectives. Finally, the researcher influences the direction of the interviews and analysis of the research. Note that no words were put into people's mouths, but interpretation is still an inherent component to the qualitative research process.

Content-wise, various limitations have been mentioned throughout the discussion already. Whilst the AI implementation life-cycle offers a nice starting point for discussion, it requires further iterations by using it in practice and verifying the assumptions (e.g. organisational, human and technical types of challenges chosen as the most relevant ones). Moreover, the operationalisation of purposes of XAI, and which types of explanations would adhere thereto, remain to be explored. Finally, zooming out to the larger scientific and societal relevance of this thesis, one of the main purposes was to overcome divides: between the 'black-box machine' and the human, between domains, and between actors. It is crucial then, that the categories and proposed multi-purpose, multi-actor subdivides, do not create further isolation, instead of serving as a shared conceptual model, or figurative coat rack, aiming at increased awareness of AI systems, including the systems' opportunities and limitations.

7.3. Directions for Further Research

This research has been approached rather holistically, with the aim of providing a comprehensible overview of AI implementation challenges and opportunities for Explainable AI to address those. A first angle for future research is to take the findings, and seek solutions or further analysis in more specialised avenues of disciplinary research, e.g. in interactive design, psychology or public administration. For instance, by seeking further investigation in branches of human-machine teaming or UX-design [121–123], but also in social sciences including looking into adoption and mental models as mentioned above [61, 118]. Another example, from a public administration or governance perspective, is integrating the AI implementation life-cycle into the public policy life-cycle, with the purpose of improving the policy operationalisation of AI-systems. The public policy cycle includes policy implementation as one of their phases; starting from agenda-setting, moving to problem-framing, decision-making, implementation and evaluation. Even though research is done on policy implementation more generally, the integration of AI there within leaves room for improvement [124]. Note that many of the ideas addressed are conceptual in nature, introducing the need to research how they would be useful in a real-world context.

A second angle for research is further probing the research findings concerning 1). the AI implementation challenges, 2). an operationalised multi-purpose, multi-actor XAI including the actor's respective information needs, and 3). the interaction between XAI and AI Implementation challenges. Ideally, one would be working towards practically usable XAI methods for every target audience, not just for the technically apt ones. One way could be to map which types of XAI, and for whom, would be most useful in every phase of the AI implementation life-cycle. Ultimately, improved insights into the AI implementation challenges and people's (information) needs improve the quality of explanations but also offer tools to face convoluted ethical considerations, prompted by working with novel techniques set within a societally-driven, yet high-impact and sensitive playing field.

8 Conclusion

The role and possible contribution of Explainable AI in the public sector, those were the focal components of this thesis at its debut. Having identified a gap between key-actors in perceiving and experiencing public sector AI implementation processes, a 'language-barrier' or difference in 'mental models' so to say, the subsequent assumption was that increasing understandability, in the form of Explainable AI from an interdisciplinary and multi-actor point of view, would help bridge a crucial discrepancy. Sixteen interviews and three observations in two Dutch public sector regulatory executive bodies were conducted, researching three main stakeholders in the AI implementation process; managers, data scientists, and domain experts. From the empirical findings, the research identified difficulty for public sector AI decision-support systems to go from a pilot phase towards deployment and adoption, even though the technical models and expertise are often in place. Also, Explainable AI would benefit from a more outspoken multi-purpose operationalisation, targeted towards different audience groups. And finally, to tackle the AI implementation challenges as currently faced by the public sector, increasing understandability is important to surmount certain challenges, yet for other challenges such as lack of priority or resources, one needs to look beyond the solution of Explainable AI.

Before diving into Explainable AI, the research explored the AI implementation challenges currently faced by the public sector cases, enabling to study the potential contribution of Explainable AI more realistically. Three types of challenges (*organisational, human and technical*) were identified, organised per phase of the AI life-cycle (*initiation & pre-conditions, design & development, pilot, deployment and evaluation*). An overview of the challenges mapped onto the AI life-cycle can be found in figure 4.1 and an overview of the AI implementation findings can be found in section 4.3. Where a majority of the challenges in the initiation phase are organisational in nature, the majority of challenges in the design and development phase are technical, and the challenges encountered in the pilot phase are predominantly human-based. As indicated in the discussion chapter, some of these mostly human-based challenges encountered during the pilot phase, such as resistance to change and lack of general knowledge about AI, might be reduced if they are recognised and handled from an earlier phase in the AI life-cycle, and by looking at concepts such as adoption mechanisms and mental models.

8.1. The Role and Possible Contribution of Explainable AI

Recall the main research question, "From a Manager's, Data scientist's and Domain expert's point of view; What is the role and possible contribution of Explainable AI for responsible implementation of algorithm-aided public decision making practices?" Explainable AI, both of the model itself as well as the reasoning behind it, could be helpful for some of these AI implementation challenges, but are often not the first things thought off. Starting with the role of XAI, a clear distinction is noticed between the two cases: the first case in which the role of Explainable AI is limited and not yet explicitly incorporated, and the other case in which they actively use and try to improve their XAI methods, by means of feature importance methods (mainly SHAP), dashboards and a review framework. The hope and focus of the managers and data scientists is to increase explainability towards the domain experts, who are the (future) users of the model. However, there is still a gap in communicating the insights, for instance of the Shapley values, to less technical audiences. Partially because the type of information does not adhere to what domain experts expect or need for their decision-making process, partially because they do not always see the value of such explanations, and partially due to more practical reasons, for instance that a dashboard to the domain expert means: yet another - currently non-integrated - information system to navigate. An overview of the findings can be found in section 5.3.

Moving to the potential contribution of Explainable AI, the research explored the information needs of the domain expert, but also the manager and data scientist. For instance, respondents would like increased or improved information about the impact of a model, about the unwanted or historical bias of the data and examples of what the model does or can do. However, the respondents needed and wanted more than just information. They also wanted increased understanding of one another, and more time or priority to incorporate AI within the organisation's current ways of working. An overview of the information and other needs can be found in figure 5.2. The information needs, help identify which type of information would be useful for the explanations, to enable increased understanding with the overarching aim of reducing the AI systems' implementation challenges.

8.2. Towards Explainable AI for the Public Sector

Explainable AI can be used most easily for the remaining technically oriented challenges in the public sector (section 6.3). For the organisational challenges, Explainable AI could provide more grounded knowledge and understanding of the AI systems at hand: to help make well-informed (political, financial and managerial) choices, to map responsibilities and understand the potential consequences of a model. However, many of the organisational challenges are not related to explanations or understanding, such as agreements, time, money, organisational structures, politics and so forth. Arguably, the real problem here is not explainability then, but other organisational challenges to be tackled simultaneously. For the human-based challenges, XAI might actually be more useful than we think; XAI has the potential to decrease resistance, and increase trust or knowledge to include ethical considerations, to name a few. Again, the understandability tools ought to be catered towards 'conversation', towards approaching the explanations from the mental models or world views as used in practice - for domain experts - and more towards accountability and rationale explanations - for managers. Also, boundaries and limitations, including positive and negative externalities of XAI deserve acknowledgement.

In the end, certain, mainly human- and organisational-, factors can simply not be addressed by elucidating a model. Some of these factors might be addressed by explanations, some of the concerns might be alleviated if understanding is enlarged. Yet, the predominantly technically oriented explanations have their limitations. Mapping the challenges in a comprehensive way, and developing a multi-purpose, multi-actor operationalised of Explainable AI is a good start to make Explainable AI more practically workable. Most importantly however, we need to prevent to use XAI just for the sake of checking a box or following a trend. Rather, it ought to really contribute to the goals we have in mind, and to society in a responsible way.

A

Appendix A: Topic list

The following topics were discussed during the interviews:

- AI implementation practices
  - The perceptions of AI
  - The perceived reasons or benefits to add AI within the organisation
  - The AI implementation process within the organisation
  - The experienced AI implementation challenges
  - The perception of other actors in the implementation process
- Explainability
  - The information needs, as desired to tackle the experienced implementation challenges
  - The perception and understanding of the concept of Explainable AI
  - The importance of XAI
  - The purposes of XAI
  - The current use of XAI
  - XAI methods
  - XAI for different target audiences
- Ideas for improvement
  - The potential for explainability to tackle implementation challenges
  - Ideas to improve XAI
  - Ideas to improve current AI implementation

Bibliography

[1] Amnesty International, *Xenofobic Machines: Discrimination through unregulated use of algorithms in the Dutch childcare benefits scandal,* (2021).

[2] J. Burrell, *How the machine 'thinks': Understanding opacity in machine learning algorithms,* Big Data and Society 3 (2016), 10.1177/2053951715622512.

[3] M. Katell, M. Young, D. Dailey, B. Herman, V. Guetler, A. Tam, C. Binz, D. Raz, and P. M. Krafft, *Toward situated interventions for algorithmic equity: lessons from the field,* undefined , 45 (2020).

[4] European Commission, *The Artificial Intelligence Act,* (2022).

[5] European Union, *Automated individual decision-making, including profiling,* (2016).

[6] K. Alfrink, I. Keller, N. Doorn, and G. Kortuem, *Tensions in transparent urban AI: designing a smart electric vehicle charge point,* 1, 3 (2021).

[7] A. Jobin, M. Ienca, and E. Vayena, *The global landscape of AI ethics guidelines,* Nature Machine Intelligence 2019 1:9 1, 389 (2019).

[8] J. Fjeld, N. Achten, H. Hilligoss, A. Nagy, and M. Srikumar, *Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-Based Approaches to Principles for AI,* SSRN Electronic Journal (2020), 10.2139/SSRN.3518482.

[9] B. Mittelstadt, *Principles alone cannot guarantee ethical AI,* Nature Machine Intelligence 2019 1:11 1, 501 (2019).

[10] A. Zuiderwijk, Y. C. Chen, and F. Salem, *Implications of the use of artificial intelligence in public governance: A systematic literature review and a research agenda,* Government Information Quarterly 38, 101577 (2021).

[11] M. Janssen, M. Hartog, R. Matheus, A. Yi Ding, and G. Kuk, *Will Algorithms Blind People? The Effect of Explainable AI and Decision-Makers' Experience on AI-supported Decision-Making in Government,* Social Science Computer Review 40, 478 (2022).

[12] T. Miller, *Explanation in artificial intelligence: Insights from the social sciences,* Artificial Intelligence 267, 1 (2019).

[13] A. Balayn, N. Rikalo, C. Lof, J. Yang, A. . Bozzon, and A. Bozzon, *How can Explainability Methods be Used to Support Bug Identification in Computer Vision Models?* (2022), 10.1145/3491102.3517474.

[14] B. Abedin, *Managing the tension between opposing effects of explainability of artificial intelligence: a contingency theory perspective,* Internet Research 32, 425 (2022).

[15] M. T. Ribeiro, S. Singh, and C. Guestrin, *"Why Should I Trust You?" Explaining the Predictions of Any Classifier,* (2016), 10.1145/2939672.2939778.

[16] K. Sokol, A. Hepburn, R. Santos-Rodriguez, and P. Flach, *bLIMEy: Surrogate Prediction Explanations Beyond LIME,* (2019).

[17] S. M. Lundberg, P. G. Allen, and S.-I. Lee, *A Unified Approach to Interpreting Model Predictions,* (2017).

[18] R. K. Mothilal, A. Sharma, and C. Tan, *Explaining machine learning classifiers through diverse counterfactual explanations,* in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (Association for Computing Machinery, Inc, 2020) pp. 607–617.

[19] D. Minh, H. X. Wang, Y. F. Li, and T. N. Nguyen, *Explainable artificial intelligence: a comprehensive review,* Artificial Intelligence Review 55, 3503 (123).

[20] OECD, *The OECD Artificial Intelligence (AI) Principles - OECD.AI,* (2019).

[21] Y. Duan, J. S. Edwards, and Y. K. Dwivedi, *Artificial intelligence for decision making in the era of Big Data – evolution, challenges and research agenda,* International Journal of Information Management 48, 63 (2019).

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

[23] C. W. Churchman, *Free for All,* https://doi.org/10.1287/mnsc.14.4.B141 14, B (1967).

[24] H. Rittel and M. Webber, *Dilemmas in a General Theory of Planning,* (1973).

[25] H. A. Simon, *Theories of bounded rationality. ,* Decision and organization 1, 161 (1972).

[26] C. E. Lindblom, *The Science of "Muddling Through",* Public administration review , 79 (1959).

[27] D. Easton, *An approach to the analysis of political systems,* (1957).

[28] C. Alden, *Critiques of the Rational Actor Model and Foreign Policy Decision Making,* in *Oxford Research Encyclopedia of Politics* (Oxford University Press, 2017).

[29] P. Rawat and J. C. Morris, *Kingdon's "Streams" Model at Thirty: Still Relevant in the 21st Century?* Politics & Policy 44, 608 (2016).

[30] M. Janssen and G. Kuk, *The challenges and limits of big data algorithms in technocratic governance,* Government Information Quarterly 33, 371 (2016).

[31] H. J. Wilson and P. R. Daugherty, *Collaborative Intelligence: Humans and AI Are Joining Forces,* Harvard Business Review 96, 114 (2018).

[32] J. Höchtl, P. Parycek, and R. Schöllhammer, *Big data in the policy cycle: Policy decision making in the digital era,* undefined 26, 147 (2016).

[33] P. Zhang, K. Zhao, and R. L. Kumar, *Impact of IT Governance and IT Capability on Firm Performance,* undefined 33, 357 (2016).

[34] M. Janssen and G. Kuk, *The challenges and limits of big data algorithms in technocratic governance,* Government Information Quarterly: an international journal of information technology management, policies, and practices 33, 371–377 (2016).

[35] A. D. Selbst, S. A. Friedler, H. College, P. A. Suresh Venkatasubramanian, and J. Vertesi, *Fairness and Abstraction in Sociotechnical Systems,* (2018), 10.1145/nnnnnnn.nnnnnnn.

[36] K. Mahroof, *A human-centric perspective exploring the readiness towards smart warehousing: The case of a large retail distribution warehouse,* International Journal of Information Management 45, 176 (2019).

[37] M. H. Jarrahi, *Artificial intelligence and the future of work: Human-AI symbiosis in organizational decision making,* Business Horizons 61, 577 (2018).

[38] T. A. Schoonderwoerd, E. M. Zoelen, K. v. d. Bosch, and M. A. Neerincx, *Design patterns for human-AI co-learning: A wizard-of-Oz evaluation in an urban-search-and-rescue task,* International Journal of Human-Computer Studies 164, 102831 (2022).

[39] ICO, *Definitions | ICO,* (2022).

[40] F. Van Harmelen and F. V. H. Nl, *A boxology of design patterns for hybrid learning and reasoning systems a preprint,* (2019).

[41] M. W. Hurley and W. A. Wallace, *Expert Systems as Decision Aids for Public Managers: An Assessment of the Technology and Prototyping as a Design Strategy,* Public Administration Review 46, 563 (1986).

[42] E. Turban and P. R. Watkins, *Integrating expert systems and decision support systems,* MIS Quarterly: Management Information Systems 10, 121 (1986).

[43] Y. Gong and M. Janssen, *An interoperable architecture and principles for implementing strategy and policy in operational processes,* Computers in Industry 64, 912 (2013).

[44] E. Brynjolfsson and T. Mitchell, *What can machine learning do? Workforce implications: Profound change is coming, but roles for humans remain,* Science 358, 1530 (2017).

[45] M. Almada, *Human intervention in automated decision-making: Toward the construction of contestable systems,* Proceedings of the 17th International Conference on Artificial Intelligence and Law, ICAIL 2019 , 2 (2019).

[46] H.-F. Cheng, R. Wang, Z. Zhang, F. O'connell, T. Gray, F. M. Harper, and H. Zhu, *Explaining Decision-Making Algorithms through UI: Strategies to Help Non-Expert Stakeholders,* in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, Vol. ACM 559, (2019).

[47] K. Kashin, G. King, and S. Soneji, *Explaining systematic bias and nontransparency in U.S. social security administration forecasts,* Political Analysis 23, 336 (2015).

[48] J. A. Kroll, J. Huey, S. Barocas, E. W. Felten, J. R. Reidenberg, D. G. Robinson, and H. Yu, *Accountable Algorithms,* 165 (2016).

[49] J. T. Browne, S. Bakker, B. Yu, P. Lloyd, and S. Ben Allouch, *Trust in Clinical AI: Expanding the Unit of Analysis,* Frontiers in Artificial Intelligence and Applications 354, 96 (2022).

[50] K. M. Scott, S. M. Wang, M. Miceli, P. Delobelle, K. Sztandar-Sztanderska, and B. Berendt, *Algorithmic Tools in Public Employment Services: Towards a Jobseeker-Centric Perspective,* ACM International Conference Proceeding Series , 2138 (2022).

[51] Y. Xie, M. Chen, D. Kao, G. Gao, and X. A. Chen, *CheXplain: Enabling Physicians to Explore and Understand Data-Driven, AI-Enabled Medical Imaging Analysis,* in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, Vol. 20 (2020) pp. 1–13.

[52] T. Q. Sun and R. Medaglia, *Mapping the challenges of Artificial Intelligence in the public sector: Evidence from public healthcare,* Government Information Quarterly 36, 368 (2019).

[53] F. Damanpour and M. Schneider, *Phases of the Adoption of Innovation in Organizations: Effects of Environment, Organization and Top Managers1,* British Journal of Management 17, 215 (2006).

[54] ICO, *An overview of the Auditing Framework for Artificial Intelligence and its core components,* (2019).

[55] K. Alfrink, I. Keller, G. Kortuem, and N. Doorn, *Contestable AI by Design: Towards a Framework,* Minds and Machines , 1 (2022).

[56] S. Fatima, K. C. Desouza, C. Buck, and E. Fielt, *Business Model Canvas to Create and Capture AI-enabled Public Value,* in *Proceedings of the 54th Hawaii International Conference on System Sciences* (2021).

[57] R. Madan and M. Ashok, *AI adoption and diffusion in public administration: A systematic literature review and future research agenda,* Government Information Quarterly 40, 101774 (2023).

[58] C. Alexopoulos, V. Diamantopoulou, Z. Lachana, Y. Charalabidis, A. Androutsopoulou, and M. A. Loutsaris, *How Machine Learning is Changing e-Government,* Proceedings of the 12th International Conference on Theory and Practice of Electronic Governance Part F148155, 354 (2019).

[59] K. C. Desouza, G. S. Dawson, and D. Chenok, *Designing, developing, and deploying artificial intelligence systems: Lessons from and for the public sector,* Business Horizons 63, 205 (2020).

[60] A. F. v. Veenstra and B. Kotterink, *Data-Driven Policy Making: The Policy Lab Approach,* (2017), 10.1007/978-3-319-64322-9_9.

[61] M. Ashok, R. Narula, and A. Martinez-Noya, *How do collaboration and investments in knowledge management affect process innovation in services?* Journal of Knowledge Management 20, 1004 (2016).

[62] E. T. Straub, *Understanding technology adoption: Theory and future directions for informal learning,* Review of Educational Research 79, 625 (2009).

[63] R. Krznaric, *The Good Ancestor* (2020).

[64] B. Goodman and S. Flaxman, *European Union Regulations on Algorithmic Decision-Making and a "Right to Explanation",* AI Magazine 38, 50 (2017).

[65] S. Bhatnagar, T. Cotton, M. Brundage, S. Avin, J. Clark, H. Toner, P. Eckersley, B. Garfinkel, A. Dafoe, P. Scharre, T. Zeitzoff, B. Filar, H. Anderson, H. Roff, G. C. Allen, J. Steinhardt, C. Flynn, S. Héigeartaigh, S. Beard, H. Belfield, S. Farquhar, C. Lyle, R. Crootof, O. Evans, M. Page, J. Bryson, R. Yampolskiy, and D. Amodei, *The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation Authors are listed in order of contribution Design Direction,* (2018).

[66] I. Calzada and C. Cobo, *Unplugging: Deconstructing the Smart City,* https://doi-org.ezproxy.library.wur.nl/10.1080/10630732.2014.971535 22, 23 (2015).

[67] J. Borrego-Díaz, J. Galán-Páez, J. Borrego-Díaz, and J. Galán-Páez, *Explainable Artificial Intelligence in Data Science From Foundational Issues Towards Socio-technical Considerations,* 32, 485 (2022).

[68] F. Caprotti, *Future cities: moving from technical to human needs,* Palgrave Communications 2018 4:1 4, 1 (2018).

[69] R. Dobbe, T. K. Gilbert, and Y. Mintz, *Hard Choices in Artificial Intelligence: Addressing Normative Uncertainty through Sociotechnical Commitments,* (2019).

[70] K. Siau and W. Wang, *Artificial Intelligence (AI) Ethics: Ethics of AI and Ethical AI,* Journal of Database Management 31, 74 (2020).

[71] A. Escobar, *Designs for the pluriverse: radical interdependence, autonomy, and the making of worlds,* Duke University Press (2018).

[72] C. Chant, *Science and technology (Chapter 3) in The SAGE Companion to the City*, edited by T. Hall, J. R. Short, and P. Hubbard (Sage, 2008) pp. 1–408.

[73] M. Muller and A. Strohmayer, *Forgetting Practices in the Data Sciences,* 19 (2022), 10.1145/3491102.3517644.

[74] S. Bankes, *Exploratory Modeling for Policy Analysis,* Operations Research 41 (1993), 10.1287/opre.41.3.435.

[75] J. H. Kwakkel, W. E. Walker, and M. Haasnoot, *Coping with the Wickedness of Public Policy Problems: Approaches for Decision Making under Deep Uncertainty,* Journal of Water Resources Planning and Management 142, 01816001 (2016).

[76] European Commission, *Assessment List for Trustworthy Artificial Intelligence (ALTAI) for self-assessment*, Tech. Rep. (2020).

[77] Ministerie van Justitie en Veiligheid, *Richtlijnen voor het toepassen van algoritmen door overheden en publieksvoorlichting over data-analyses.* (2021).

[78] Gemeente Amsterdam, *Gebruik van algoritmes binnen gemeente Amsterdam,* (2023).

[79] A. S. Franzke, I. Muis, and M. T. Schäfer, *Data Ethics Decision Aid (DEDA): a dialogical framework for ethical inquiry of AI and data projects in the Netherlands,* Ethics and Information Technology 23, 551 (2021).

[80] I. Georgieva, C. Lazo, T. Timan, and A. Fleur Van Veenstra, *From AI ethics principles to data science practice: a reflection and a gap analysis based on recent frameworks and practical experience,* AI and Ethics 2021 2:4 2, 697 (2022).

[81] European Commission, *Developments | The Artificial Intelligence Act,* (2023).

[82] J. C. Heilinger, *The Ethics of AI Ethics. A Constructive Critique,* Philosophy and Technology 35, 1 (2022).

[83] M. VEALE, *A Critical Take on the Policy Recommendations of the EU High-Level Expert Group on Artificial Intelligence,* European Journal of Risk Regulation 11, e1 (2020).

[84] S. Larsson, *On the Governance of Artificial Intelligence through Ethics Guidelines,* Asian Journal of Law and Society 7, 437 (2020).

[85] A. Matthias, *The responsibility gap: Ascribing responsibility for the actions of learning automata,* Ethics and Information Technology 6, 175 (2004).

[86] F. Santoni de Sio and G. Mecacci, *Four Responsibility Gaps with Artificial Intelligence: Why they Matter and How to Address them,* Philosophy and Technology 34, 1057 (2021).

[87] J. Gardner, *The Mark of Responsibility,* Offences and Defences , 177 (2007).

[88] M. Bovens, *Analysing and Assessing Accountability: A Conceptual Framework1,* European Law Journal 13, 447 (2007).

[89] M. Hildebrandt, *Privacy as protection of the incomputable self: From agnostic to agonistic machine learning,* Theoretical Inquiries in Law 20, 83 (2019).

[90] G. Noto la Diega and L. Diega, *Against the Dehumanisation of Decision-Making Algorithmic Decisions at the Crossroads of Intellectual Property, Data Protection, and Freedom of Information,* (2018).

[91] T. Hirsch, K. Merced, S. Narayanan, Z. E. Imel, and D. C. Atkins, *Designing Contestability,* in *Proceedings of the 2017 Conference on Designing Interactive Systems* (ACM, New York, NY, USA, 2017) pp. 95–99.

[92] T. Hagendorff, *The Ethics of AI Ethics: An Evaluation of Guidelines,* Minds and Machines 30, 99 (2020).

[93] European Commission, *Ethics guidelines for trustworthy AI | Shaping Europe's digital future*, Tech. Rep. (2019).

[94] A. A. Tubella, A. Theodorou, V. Dignum, and F. Dignum, *Governance by Glass-Box: Implementing Transparent Moral Bounds for AI Behaviour,* (2019).

[95] A. Barredo Arrieta, N. Díaz-Rodríguez, J. D. Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garcia, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila, and F. Herrera, *Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI,* Information Fusion 58, 82 (2020).

[96] R. Dazeley, P. Vamplew, C. Foale, C. Young, S. Aryal, and F. Cruz, *Levels of explainable artificial intelligence for human-aligned conversational explanations,* Artificial Intelligence 299, 103525 (2021).

[97] Cambridge Dictionary, *Explanation | English meaning - Cambridge Dictionary,* (2023).

[98] A. Adadi and M. Berrada, *Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI),* 10.1109/ACCESS.2018.2870052.

[99] Y. K. Dwivedi, N. P. Rana, M. Janssen, B. Lal, M. D. Williams, and M. Clement, *An empirical validation of a unified model of electronic government adoption (UMEGA),* Government Information Quarterly: an international journal of information technology management, policies, and practices 34, 211 (2017).

[100] R. S. Verhagen, S. Mehrotra, M. A. Neerincx, C. M. Jonker, and M. L. Tielman, *Exploring Effectiveness of Explanations for Appropriate Trust: Lessons from Cognitive Psychology,* (2022).

[101] S. Raghunathan, *Impact of information quality and decision-maker quality on decision quality: a theoretical model and simulation analysis,* Decision Support Systems 26, 275 (1999).

[102] A. Rai, *Explainable AI: from black box to glass box,* (2019), 10.1007/s11747-019-00710-5.

[103] M. Nassar, K. Salah, M. Habib ur Rehman, and D. Svetinovic, *Blockchain for explainable and trustworthy artificial intelligence Technologies > Machine Learning Technologies > Computer Architectures for Data Mining Fundamental Concepts of Data and Knowledge > Key Design Issues in Data Mining,* (2019), 10.1002/widm.1340.

[104] D. Slack, S. Hilgard, E. Jia, S. Singh, and H. Lakkaraju, *Fooling LIME and SHAP: Adversarial Attacks on Post hoc Explanation Methods,* (2020), 10.1145/3375627.3375830.

[105] J. van der Waa, E. Nieuwburg, A. Cremers, and M. Neerincx, *Evaluating XAI: A comparison of rule-based and example-based explanations,* Artificial Intelligence 291 (2021), 10.1016/J.ARTINT.2020.103404.

[106] S. Wachter, B. Mittelstadt, and C. Russell, *Counterfactual explanations without opening the black box: automated decisions and the GDPR,* (2017).

[107] M. T. Ribeiro, S. Singh, and C. Guestrin, *Anchors: High-Precision Model-Agnostic Explanations,* Proceedings of the AAAI Conference on Artificial Intelligence 32, 1527 (2018).

[108] ICO, *Explaining decisions made with AI: What goes into an explanation?* (2022).

[109] D. Pedreschi, F. Giannotti, R. Guidotti, A. Monreale, S. Ruggieri, and F. Turini, *Meaningful Explanations of Black Box AI Decision Systems,* Proceedings of the AAAI Conference on Artificial Intelligence 33, 9780 (2019).

[110] H. de Bruijn, M. Warnier, and M. Janssen, *The perils and pitfalls of explainable AI: Strategies for explaining algorithmic decision-making,* Government Information Quarterly 39 (2022), 10.1016/j.giq.2021.101666.

[111] S. Robbins, S. Robbins scott, and S. Robbins, *A Misdirected Principle with a Catch: Explicability for AI,* Minds and Machines 2019 29:4 29, 495 (2019).

[112] G. D'Acquisto, *On conflicts between ethical and logical principles in artificial intelligence,* AI & SOCIETY 2020 35:4 35, 895 (2020).

[113] M. Foucault, *Discipline and Punish : the Birth of the Prison.*, Tech. Rep. (Pantheon Books,, New York, 1977).

[114] V. Braun and V. Clarke, *Feminism & Psychology*, 3 (2013).

[115] R. Yin, *Case study research: Design and methods,* Sage 5 (2009).

[116] C. de Haas, H. Kaya, and A. Qahtan, *Usability Study of an Explainable Machine Learning Risk Model for Predicting Illegal Shipbreaking,* Utrecht University (2021).

[117] S. Treur, *Designing an Interface for an Explainable Machine Learning Risk Model for Predicting Illegal Shipbreaking,* Utrecht University (2022).

[118] H. Rutjes, M. C. Willemsen, and W. A. IJsselsteijn, *Considerations on explainable AI and users' mental models,* (2019).

[119] S. Mohseni, N. Zarei, and E. D. Ragan, *A Multidisciplinary Survey and Framework for Design and Evaluation of Explainable AI Systems,* ACM Transactions on Interactive Intelligent Systems (TiiS) 11 (2021), 10.1145/3387166.

[120] L. Carlsson, E. A. Shanahan, M. K. Mcbeth, and P. L. Hathaway, *Policy Science at an Impasse: A Matter of Conceptual Stretching?* Politics & Policy 45, 148 (2017).

[121] Q. V. Liao, D. Gruen, and S. Miller, *Questioning the AI: Informing Design Practices for Explainable AI User Experiences,* Conference on Human Factors in Computing Systems - Proceedings (2020), 10.1145/3313831.3376590.

[122] H. Subramonyam, J. Im, C. Seifert, and E. Adar, *Solving Separation-of-Concerns Problems in Collaborative Design of Human-AI Systems through Leaky Abstractions,* Conference on Human Factors in Computing Systems - Proceedings (2022), 10.1145/3491102.3517537.

[123] W. Jin, J. Fan, D. Gromala, P. Pasquier, and G. Hamarneh, *EUCA: the End-User-Centered Explainable AI Framework,* Proceedings of 1 (2021).

[124] D. Valle-Cruz, J. I. Criado, R. Sandoval-Almazán, and E. A. Ruvalcaba-Gomez, *Assessing the public policy-cycle framework in the age of artificial intelligence: From agenda-setting to policy evaluation,* Government Information Quarterly 37, 101509 (2020).