Towards Responsible AI Adoption: Implementation Challenges in the Dutch Public Sector from a Multi-Actor Perspective

ANONYMOUS AUTHOR(S)

To foster a culture of responsible use of AI in public decision-making, current design and implementation practices and challenges should be researched and shared. Yet, little is discussed about how AI projects are implemented in practice, less so in the public sector. In this paper, we look at AI implementation processes within two Dutch executive agencies. We conducted 16 semi-structured interviews with data scientists, project managers and decision-makers, and observed three inter-actor project meetings. We found that the lack of shared understanding around the use of AI, difficulties in integrating AI into existing processes (e.g., while data scientists and managers are encouraged to implement AI projects, decision-makers do not always see the added value to their work), and lack of priority and resources (either for decision-makers to contribute to the AI projects, or to spend time to educate and responsibly introduce AI models by data scientists), require the need to incorporate organizational preconditions. This includes investing in a common language between AI and domain knowledge, and by deliberating the algorithmic decision-making systems' underlying values. Moreover, the sensitive political context, ethical concerns, and unclear responsibility attribution pose challenges to AI implementation projects. Based on these insights, we suggest to enhance value-driven transparency, and to provide explanations that allow for discretion and contestability by the key actors involved in the AI implementation process, such that AI adoption in the public sector is warranted and meaningful.

CCS Concepts: • **Human-centered computing → Empirical studies in HCI**; **Collaborative and social computing**; • **Computing methodologies → Machine learning**.

Additional Key Words and Phrases: artificial intelligence, machine learning, algorithmic decision-making, public administration, implementation challenges

ACM Reference Format:
Anonymous Author(s). 2018. Towards Responsible AI Adoption: Implementation Challenges in the Dutch Public Sector from a Multi-Actor Perspective. 1, 1 (December 2018), 28 pages. https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

Where Artificial Intelligence (AI) [1] has become an ubiquitously used term, the public sector seeks to keep abreast. A few examples of algorithmic systems for public services include: policing [11], public employment and medical services [91, 96, 111], unemployment benefits [92] and access to

[1] Throughout this paper, we will use the term AI to refer to computational systems that are designed for interpreting external data, learning from that data, and using those learnings for performing specific tasks [49]. We will use the term *algorithmic decision-making* to refer to decision-making processes that are driven or augmented by AI systems. It should be noted that these are the *working definitions* used for the sake of clarity *in the paper*. For the interviews, we did not define AI to our interviewees since we were interested in knowing how the interviewees conceptualized and described AI in their own words (see section 5.2.1).

Author's address: Anonymous Author(s).

XXXX-XXXX/2018/12-ART $15.00
https://doi.org/10.1145/1122445.1122456

housing support [98]. Despite claims about the potential of AI systems to increase efficiency, lower costs [52], and provide better outcomes for citizens [86] by informing policy making [52], when these systems are not implemented[2] and used responsibly, they can have adverse consequences, and give rise to, for instance, discrimination based on gender or race [27, 35, 94]. By now, most researchers and practitioners agree that algorithmic systems used in the public domain need to adhere to a minimum set of Responsible AI values, such as "Accountability", "Explainability"[3] "Fairness", and "Contestability" [32, 48]. The difficulty arises when translating these values into organizational and technical workflows [72].

Responsible AI implementation and use in the public sector faces several practical challenges stemming from two main co-occurring sources of complexity: (1) the nature of AI itself [99, 112], and (2) the unique conditions under which such systems must operate within the public sector [90]. Due to its demand for data, technical complexity, and unpredictable interactions [112], AI and human-AI interactions are uniquely difficult to design for [18, 99, 112]. The application of AI in the public sector exacerbates the impact of such issues, where societally sensitive topics and a multi-actor political playing field are the norm. Algorithmic decision-making in the public sector has been claimed to need a distinctly unique framework [90] to account for the complexities of the bureaucratic processes [58] under which these systems must operate and to adequately afford human discretion [6] once they are deployed. We echo these claims and argue that AI systems used for decision-making within the public sector need to be carefully scrutinized.

Despite recent advances in the scrutiny of public AI systems, there is still little *empirical* insight into the challenges of responsibly *implementing* AI in the public sector. Previous work mainly focused on the identification of challenges that arise whilst *using* AI systems, or suggested *theoretical* design interventions and guidelines to address such challenges (e.g., [12, 36, 43, 90]). However, such interventions and guidelines can only be effective if they consider and align with the processes that lead to the development and implementation of AI systems in the first place [64]. One of those design guidelines, for example, suggests that algorithmic outputs in the public sector should be multidimensional rather than singular metrics so that decision-makers can exercise discretion [90]. To answer to previous research where collaboration difficulties have been detected, and the early involvement of decision-makers has been vouched for, yet empirical research tends to focus on one actor at a time [50, 90], we adopt a multi-actor perspective.

In this paper, we argue that exploring fundamental challenges that arise during the implementation of AI projects offers an opportunity to understand how and why public AI systems may lead to harmful downstream consequences. More specifically, understanding these processes as perceived by the different key-actors, will offer a layer of in-depth analysis; identifying the commonalities and tensions between their practices and the challenges they face. For instance, if we had only interviewed data scientists, we would have known that they are missing the input from decision-makers. By also incorporating the decision-makers, we learned why they do not always cooperate (e.g., lack of time, ethical concerns), and by including the managers, we learn how such decisions are made and/or cannot always be influenced by the actors themselves. To this end,

[2] The process of technology *adoption* in organizations has been defined as the systematic approach of integrating a technology in an organization so that it is appropriately used by all concerned actors [88]. Based on this theoretical conceptualization, we will use the term *implementation* to refer to the events and actions that pertain to preparing the organization for the use of an AI system, its trial use, and acceptance by the users [24]. When referring to previous work on the *use* of AI systems, we point to the continued employment of AI systems as part of the routine of an organization.

[3] The initial research question focused on explainability/Explainable AI (XAI) practices in the public sector. However, the first interviews clearly indicated that such AI functionality [81] was hardly present. Therefore, we sought a more fundamental understanding of AI implementation, before diving into XAI as a potential solution.

we focus on AI projects taking place in two executive agencies[4] of the Dutch public sector: the Human Environment and Transport Inspectorate (*De Inspectie Leefomgeving en Transport or ILT*), and another agency we need to refer to anonymously due to confidentiality reasons[5]. The Dutch public sector represents a relevant context to study AI implementation projects due to three main reasons. First, the Netherlands pioneers in experimenting with the use of AI technologies for public services at a European level [70, 105]. Second, recent scandals due to the inappropriate use of AI for public services raised the Dutch public sector to international notoriety [8]. The national childcare benefits scandal or '*de toeslagenaffaire*'[6] greatly contributed to bringing the issue to light. And third, the increasing commitment of the Dutch public sector towards enabling scrutiny of the AI systems they use[7]. It is this relatively normalized culture of openness to scrutiny that allowed us, as researchers, to get access to our case studies, and to be able to publish our findings, yet under confidentiality constraints.

Our work, therefore, offers the CSCW community an opportunity to get first-hand real-world accounts from different actors in a context where the usage of AI has been embraced, as well as criticized by its downside effects. From a multi-actor (i.e., involving managers, data scientists, and decision-makers[8]) perspective, we seek to answer the following research question:

> **What are the challenges of implementing AI in the Dutch public sector, as experienced by managers, data scientists and decision-makers?**

Between October 2022 and January 2023 we interviewed 5 managers, 7 data scientists, and 4 decision-makers. Additionally, we conducted two observations of multi-actor project meetings and one observation of decision-makers' day-to-day workflows. Our results show that, in the selected use cases, AI is introduced by a dedicated innovation team. Despite acknowledging the promises of AI, there is still a lack of shared understanding among the key actors on what AI is, and why or how it should be used. Various AI projects are being implemented, yet the actors experience difficulties in adopting AI within the existing organizations' workflows (e.g., while data scientists and managers are encouraged to implement AI projects, decision-makers do not always see the added value to their work). The actors express a lack of priority and resources; either to contribute to the AI projects (decision-makers) or to spend time to educate and responsibly implement AI projects (data scientists). Heightened by the Dutch politically sensitive playing field, and due to prior negative experiences with the large-scale use of AI in the Dutch public sector, political scrutiny is referred to as a challenge by all three actor-groups, though with the nuance that managers refer more to larger questions of public accountability, and decision-makers are worried that their discretionary power might be compromised. Ethical concerns are raised; with fairness considered mainly by managers and data scientists, and concerns about trust and explainability highlighted by decision-makers. Finally, all actors signal a lack of responsibility attribution, where the responsibilities – especially when it comes to accounting for ethical concerns – within AI projects are not clearly defined. With

[4] By executive agency, we refer to the Dutch concept of 'Uitvoeringsorganisatie'. These are national agencies delivering public services, often characterized by having to implement complex laws and regulation in a fast-changing society. For many citizens, these agencies are their direct connection to the government [14, 84].

[5] Note that both cases have politically sensitive workflows, therefore, traceable information about the workings of the algorithms or the persons involved cannot be disclosed.

[6] The childcare benefits scandal entailed that tens of thousands of parents and caregivers were falsely accused of childcare benefit fraud by the Dutch tax authorities.

[7] https://algoritmes.overheid.nl/nl/algoritme (last accessed 02.07.2024).

[8] Decision-makers are public servants that use, or will use, algorithmic outputs for their decisions. They are also the "domain experts" and "end users" of the algorithmic systems.

this study we contribute novel insights into the practical challenges arising from the process of designing and implementing responsible AI solutions in the public sector. We do so, by taking a step back from the use of AI systems, and by identifying challenges that arise as early as the conception and implementation of public sector algorithmic decision-making systems.

2 RELATED WORK

This section first describes previous research on principles and guidelines for Responsible AI and their (mis)alignment with practice (section 2.1). Next, we synthesize related work on the use of AI in the public sector, and the challenges to implement AI responsibly (section 2.2).

2.1 Researching Responsible Practices in AI Development and Implementation

In an effort to ensure the ethical development and use of AI, organizations both in the public [19, 30, 31, 48, 77, 100] and the private sphere [38, 45, 67] have published documents and policy efforts to guide the ethical development and implementation of algorithmic decision-making systems[48]. Yet, these may be too high level to operationalize [72, 75, 116].

Recent studies, mostly conducted in private companies (e.g., [41, 62, 82]), have shown that applying these ethical principles is not straightforward and that practitioners struggle when putting responsible AI guidelines into practice. While AI practitioners try to learn about responsible AI through e.g., information foraging or interpersonal learning, many of the available responsible AI resources are limited to computational approaches [63]. For example, for the principle of *fairness*, toolkits have been developed [42]. Yet, whilst looking into *fairness* in practice, studies such as conducted by Deng et al. [25, 26] have shown that these toolkits are often driven by availability of algorithmic methods, rather than by real-world needs, and that AI practitioners need expert guidance to be able to effectively make choices in their fairness analysis. Previous studies have also shown that the use of fairness toolkits may devolve into a checkbox culture rather than a practice [9]. Beyond *fairness*, for the principle of *transparency*, similar examples of discrepancies between theory, methods and practice can be found. Heger et al. [41] researched the usage of data documentation artefacts (i.e., *Datasheets* [37]), claiming that AI practitioners found these artefacts to be difficult to integrate into their existing workflows.

The aforementioned works demonstrate that there is often a mismatch between the theoretical guidelines for the responsible development and implementation of algorithmic decision-making systems and the reality on the ground. In our study, we conduct interviews with actual practitioners (i.e., project managers, data scientists, and decision-makers) that are currently working on, or have recently experienced, the implementation of algorithmic decision-making systems. We consider these first-hand testimonies as an initial necessary step towards suggesting design choices to further promote responsible practices in AI implementation.

2.2 Challenges to Responsible AI in the Public Sector

AI has been claimed to have enormous benefits for the public sector [86]. In a sector that is uniquely complex and requires sophisticated methods [33] to navigate bureaucracy [90]. AI is believed to enable increased efficiency, lower costs [52] and provide individual citizens better outcomes [86] by processing big amounts of data in a more consistent way than humans [52]. As a result, algorithms can be found in various areas of the public domain, e.g; in the detection of fraud [117], public employment and medical services [91, 96, 111], the development of climate policies [103], and policing [11, 66] [9]. Despite the increased popularity of applying AI to public sector practices, prior

[9] Engin and Treleaven [28] developed a taxonomy for algorithmic governance; this paper is mainly concerned with their category of "supporting civil servants" with algorithmic decision-making.

work highlighted several challenges that arise due to the specific context of implementation; the public sector context.

The public sector deals with high-impact, complex and societally relevant topics. Underlying many of these public sector specific challenges, are issues related to legality and often competing public values. Consequently, some challenges are given special attention in public administration literature, among which; unpreparedness for a fundamental change in decision-making processes introduced by AI, underestimating the human role in designing AI systems, and (not) allowing for discretionary power given non-flexible algorithmic behaviour [93].

2.2.1 *Unpreparedness for Fundamental Organizational Changes Introduced by AI.*

Introducing AI into public decision-making, implies a fundamental shift in how public organizations function and in the role of public managers [110], yet they are still unprepared for the challenges introduced by working with AI [1]. Since the nature of decision-making in public administration is particular – bureaucratic, inherently political, time-consuming and conflict-invoking [40] – there has been a long tradition of literature focusing on public decision-making practices (e.g. [20, 87]) before the introduction of AI. Scholars emphasize that generally, public decisions“need to be solved based on incomplete, contradictory, and changing information” [47]. Selten and Meijer [93] analyze how literature often considers that AI builds or erodes public values [10]. Based on Hood [44]’s framework, Selten and Meijer [93] categorize three types of public values; sigma values (values related to efficiency and effectiveness), theta values (values related to fairness and transparency), and lambda values (the ability of organizations to be adaptive and robust). Where the first set of values is often positively influenced by AI, and the second negatively, the third is underexposed in the current academic debate [93]. The values are not always compatible with one another; citizens may expect efficiency, transparency and robustness, but they might not be feasible at the same time. Governing the complexity of public algorithmic decision-making will require trade-offs; posing an important challenge for public decision-makers [93].

Valle-Cruz et al. [103] argue that for AI implementation in the public sector to succeed, we need to ensure that (1) it does not conflict with operational requirements and routines, (2) AI is not perceived as inappropriate by operational experts for the tasks of the organization. A third point can be added, based on the work of Selten and Meijer [93], where (3) public values and their trade-offs need to be considered.

2.2.2 *Underestimating Complexities in Human-AI Interactions during Design.*

Designing an AI system determines how people might be able to use the system. Yang et al. [113] map various human-AI interaction design challenges, e.g., difficulty to articulate what AI can/cannot do, the technical feasibility being highly dependent on the available data, not knowing how to purposefully use AI in the design problem at hand. Furthermore, AI has a political dimension, with the ability to confirm or oppose prevailing opinions based on so-called objective facts and numbers [93]. Alon-Barkat and Busuioc [7] look into public sector human-AI interactions, identifying two important AI-related biases from literature. The first is automation bias; where the algorithmic advice is followed even if other sources exhibit “warning signals” [7]. The second is selective adherence, where people selectively adopt algorithmic advice corresponding to stereotypes [7]. While their research does not confirm automation bias, it does find evidence for selective adherence.

Meijer et al. [65] identify two emerging patterns of interacting with AI systems in a public sector context – ‘the algorithmic cage’ and ‘the algorithmic colleague’. The algorithmic cage surfaces in hierarchical administrative cultures, where organizational power is exerted through the AI system.

[10] Public values here are defined as “citizens collective expectations in respect to government and public services” [74].

The algorithmic colleague in contrast, provides sensible advice considered vis-a-vis contextual knowledge. They argue that in the Netherlands one is more likely to find the algorithmic colleague, because of their support of professional judgment and discretion compared to other countries, in this case Germany. However, the diminishing and shifting of human discretion due to the introduction of AI, which also concerns the Dutch public sector, is emphasized and disputed by other authors (elaborated upon in next subsection 2.2.3), highlighting the necessity to better crystallize the human role in algorithmic decision-making systems. These different emerging patterns illustrate the complexity inherent to designing human-AI interactions, due to their unpredictability [113].

2.2.3 Shifting Discretionary Power due to the Arrival of AI. Having a certain amount of room for interpretation and discretion in decision making is vital for the work of civil servants [6, 55, 76]. AI systems used in the public sector need to afford street-level bureaucracy and allow public administrators to exercise discretion [90]. Consequently, they should enable street-level public administrators to apply their tacit and explicit knowledge for making complex decisions and to ensure that human and democratic values are considered [43]. However, the discretionary powers of the street-level professionals in public organizations have been altered by AI systems [6, 76, 93, 117]. Alkhatib and Bernstein [6] introduce the concept of street-level algorithms: where *street–level bureaucrats* reflexively refine their decision criteria by reasoning through a new situation, *street–level algorithms*, at best, refine their criteria after the decision is made [6].

Zouridis et al. [117] recognize a shift in discretionary power due to the introduction of public algorithmic decision-making systems; from decision-makers to those programming the algorithms and the data analysts. The locus of administrative discretion has shifted to those responsible for programming the decision-making process and translating the legislation into software [117]. Public agencies often acquire such systems through government procurement processes, resulting in the delegation of software design choices to third-party developers and allowing little public participation in decisions about goals or values [76]. In the case of data-driven AI systems particularly, it is not even the analysts of the data that have the discretionary power, but rather the ones programming the algorithms, which as opposed to traditional rule-based algorithms are not hard coded [117]. Where some authors focus on the shift in discretionary power, others emphasize the decrease or loss thereof, yet all emphasize the importance of designing for- and ensuring decision-making discretion [6, 16, 76, 93, 117].

In this paper, we explore the implementation of AI projects in the public sector. We echo Madan and Ashok [64] who vouch to approach this "*through* ***in-depth*** *case studies or* ***ethnographic*** *studies outlining the underlying mechanisms and dynamics of AI projects*". By focusing on a thorough diagnosis of these underlying mechanisms and dynamics in the Dutch public sector— specifically including the perspectives of non-technical as well as technical actors — we 1). offer a holistic view of public sector-specific challenges to implementing AI in a responsible manner, 2). dissect how the challenges previously found (described in section 2.2), surface during implementation of AI in a practical setting, and 3). outline these processes as perceived by different key-actors. While acknowledging that some of these practices and challenges are contextual and situated in nature, we aim to propose first steps towards designing AI implementation processes that can mitigate or prevent harmful algorithmic consequences in a public sector environment.

3 BACKGROUND: AI IN THE DUTCH PUBLIC SECTOR

In this section, we give insight into the nature of the Dutch public sector. The Dutch public sector represents a relevant context to study the dynamics in AI implementation projects for three main reasons: (1) the increasing use of AI for public services, (2) recent scandals due to the inappropriate use of AI, and (3) initiatives towards improving scrutiny of such systems.

Increasing use of AI. The Netherlands has a strong tradition of incorporating technical tools (e.g., computer simulations for environmental assessments [79], planning support systems [109]) in decision-making processes. Artificial Intelligence systems have been no exception. In 2018, AI became an official independent policy domain in the Netherlands [106]. It was designated the status of a "key technology" for societal and economic chances [83]. Initial mappings of AI technology for public services in Europe, found that the highest amount of initiatives to use AI-technologies was in the Netherlands [70, 105]. AI systems supporting public decision-making in the Netherlands 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 [47]. However, Janssen et al. [47] 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 machine learning algorithms, their benefits and challenges are less well understood [47].

Recent Scandals. In recent years, the (ir)responsible use of AI in public services has been high on the political agenda, most prominently following the Dutch childcare benefits scandal [8]. Brought to public attention in 2018, the scandal led to the fall of the national government in 2021 and its effects are ongoing in the Dutch political- and social landscape today: ranging from news articles, policy implications, ongoing political debates, international attention, and, critique on using algorithms for delicate topics such as allocating childcare support[11]. Other scandals such as the digital welfare fraud detection system called *Systeem Risico Indicatie* (SyRI) [80, 104] reinforced the critique and attention drawn to the application of AI systems in Dutch public decision-making. However, simultaneously, the increasing push for digitization and the use of algorithms [47], exposes a tense playing field with conflicting demands.

Initiatives to Improve Scrutiny. The Dutch government is making an effort to enhance scrutiny of AI systems, by launching an algorithm register where local governments can publish their algorithms [12]. The goal behind algorithm registers is to make AI solutions used for public services as *"responsible, transparent, and secure"* [34] as possible. At the time of writing this manuscript (July 2024), there are 438 entries, each including general information about the name of the AI system, organization that is using or developing it and a short description (among other details). While the algorithm register initiative has been commended for representing a step towards making public AI systems open to scrutiny, it has also received several criticisms due to the voluntary participation to its inventory; and to the insufficient level of details provided by its current entries.

4 METHODOLOGY

A mixed-method qualitative approach is adopted to perform this research. Given the various perspectives and interpretations of stakeholders involved, it suits to do in-depth analysis of those visions [21]; building from interviews and observations.

[11] A few examples are: (a) https://www.nrc.nl/nieuws/2023/10/09/inspectie-hersteloperatie-schaadt-ernstig-gedupeerden-toeslagenschandaal-a4176601; (b) https://www.theguardian.com/world/2021/jan/14/dutch-government-faces-collapse-over-child-benefits-scandal; (c) https://www.tweedekamer.nl/debat_en_vergadering/commissievergaderingen/details?id=2023A00303 (last accessed 02.07.2024).

[12] https://www.digitaleoverheid.nl/overzicht-van-alle-onderwerpen/algoritmes/algoritmeregister/ (last accessed 02.07.2024).

4.1 Case Study Design

The research uses a multiple case study design. Within the Dutch public sector, we chose two case studies: the Human Environment and Transport Inspectorate (*De Inspectie Leefomgeving en Transport* or ILT), and a second case study that we are not able to disclose due to confidentiality requirements. We chose these cases because (1) they are prominent in the Dutch public sector, therewith serving as elucidating examples, (2) we wanted to study public bodies operating on a national scale, and (3) both cases have an executive- and supervisory character (e.g., the tax authority or the police) rather than a policy-making character (e.g., a ministry). In the Netherlands, these executive agencies are called 'Uitvoeringsorganisaties'. Executive agencies are responsible for translating policy to practice [5, 14], therewith being *executive* and *not legislative* in nature. The chosen organizations are among 38 prominent executive agencies, studied for their large-scale contact with citizens and businesses in the 2022 annual report on the "state of public service delivery" [84]. The studied organizations have the aim of integrating AI in their core activities and workflows, using algorithmic decision-making to assist civil servants in their supervisory tasks. Both organizations have previously experimented with AI systems, though present different levels of AI maturity; the ILT having more experience with different algorithmic decision-making systems, explained in more detail below. Especially since the government as 'user' of AI technologies has received less attention in literature than the government as 'regulator', these executive agencies are interesting organizations to study [53, 118].

The ILT is one of the 11 official inspectorates of the Netherlands [85]. It is the supervising agency of the Ministry of Infrastructure and Water Management, with 1600 employees, and its focus is to maintain the safety, trust and sustainability in the transportation system, infrastructure, and the living environment of the Dutch Society [29]. It does so by monitoring, checking and supervising more than 170 laws related to multiple topics, such as transportation of dangerous goods via air, sea and water, international shipment of waste, plastic processing or quality of the rail infrastructure. Due to the multitude of topics, and the amount of units to be inspected for every topic, the ILT must be both selective, and effective. Selective because deployment is focused on high risk targets and effective because officials are provided with the means to act. These objectives are believed to be better achieved with a data-driven approach, and The Innovation and Data Lab of the ILT (IDlab) was given the assignment to support the organization in this transition. To this end, the IDlab tests and develops Trustworthy AI models on a multitude of topics and data types, while following the national legislation and EU AI Act [13]. The second case study is also a national public executive agency (*uitvoeringsorganisatie*) in the Dutch public sector, enforcing politically sensitive policies, and increasingly exploring and introducing AI to support their decision-making processes. Due to confidentiality, we cannot give a detailed description about the role of AI in that executive agency. We can, however, mention that, similar to ILT, AI is being implemented with the hope of making public decision-making more efficient, and complementing the expertise of decision-makers with data-driven insights.

4.2 Semi-Structured Interviews with Civil Servants

Interviews form the main source of insights and empirical results. 16 semi-structured interviews were conducted by the first author, divided over the two cases; 5 managers, 7 data scientists and 4 decision-makers. Each respondent is a civil servant employed by the respective organizations. For the interviews, they were recruited per defined role. The role of *manager* refers to the actors deciding whether and how to include AI on a strategic level. The managers were responsible for a

[13] Note that the ILT does not plan to implement fully autonomous AI systems. The inspectors ultimately make the final decision.

variety of topics, but always including one or more AI projects, and decisions to be made about AI implementation and use. *Data scientist* refers to those developing and validating the algorithms, and *decision-maker* to those using, or potentially using, the algorithms in their workflows, i.e., they are also the 'end-users' and 'domain experts'. Each interview was conducted individually, with an average duration of 60 minutes. Our study was approved by the ethics committee at our research institution, and both cases consented to publish this work. We do not separately report the insights of each use case to ensure confidentiality and to protect the privacy of the participants involved. The semi-structured interviews were conducted based on a topic list, to be be found in Appendix A. The semi-structured nature of the interviews allowed for discussion of the listed topics, while leaving room for respondents to elaborate on unanticipated issues [21].

Observing three gatherings with key actors regarding their AI implementation processes, added to the insights from the interviews[14]. Due to the sensitive nature of topics, the meetings were not recorded nor transcribed. Instead, the first author made detailed minutes of the meetings, later analyzing these with a similar protocol as the transcribed interviews. While recordings might have been desirable from a methodological point of view, they could have interfered with openness and frankness of the discussions, in the given politically-charged contexts [107].

4.3 Data Analysis

We analyzed the bluedata (i.e., both the interviews and the minutes of the meetings) using *reflexive thematic analysis* [22], which is a fitting method for interviews and observations; allowing to adapt for a range of questions, observations, and categorize the answers provided by the respondents [21]. The procedure to go from interviews and observations to results and interpretation was: (1) *Transcribing the interviews*[15]. (2) *Coding the interviews and observations* using the software ATLAS.ti[16]. Selective coding was applied to more general topics identified in the literature study (e.g., the AI implementation process), whereas open coding was applied to the concepts that were more sensitive to empirical experiences and interpretations (e.g., AI implementation challenges). (3) *Reviewing and mapping the themes* per category. Every code was considered and divided into organizational, technical, and socio-technical challenges. This served as a first mapping. Since strictly technical challenges were hardly mentioned these are not reflected as a separate sub-category in the final results. (4) *Defining overarching themes, AI implementation challenges and additional insights.* We clustered the code groups to define overarching themes in the final results section. Some code groups were excluded, due to confidentiality reasons. Decisions to cluster or exclude codes were never done by one researcher, but always after deliberation with the paper's authors. To validate the interpretations, the results were presented to and discussed with a sample of the respondents.

4.4 Positionality

To position the findings, it is essential to acknowledge that people's experiences and worldviews are part of qualitative research and meaning-making within a real-life context [21]. This holds true both for the respondents and the researchers. With a background in (participatory) governance as well as software development in the Dutch public sector, the interviewing researcher shared cross-disciplinary knowledge with the different actors, making it easier to interact with them.

[14] Political sensitivity and busy schedules reduced the access especially to decision-makers. The respondents in one of the cases outnumbered those in the other, but having access to the documents around the organization's AI implementation process and observing meetings still allowed in-depth insights.

[15] Most interviews were conducted in Dutch, and all direct quotes from respondents were translated from Dutch into English by the first author.

[16] https://atlasti.com/.

Furthermore, all associated researchers contribute a mix of disciplinary backgrounds, including in the domains of human-centered AI, AI ethics and responsible AI.

5 RESULTS

This section analyzes AI implementation challenges, as perceived by key actors. 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*" (DS7). Rather than technical challenges, political-, organizational- and socio-technical challenges (section 5.2) were mostly highlighted throughout the interviews. A categorization of our results is provided in Table 1. We chose to focus on the challenges that had the most apparent commonalities and differences among stakeholders, and are most specific to the public sector context. Respondents include Manager (M_i), Data Scientist (DS_j), Decision-Maker (DM_k)[17]. Additionally, three observations (OB_l) are incorporated.

Multi Actor Challenges in AI Implementation	
5.2.1	Lack of Shared Understanding
5.2.2	Difficulties to Integrate AI in Current Workflows
5.2.3	Lack of Priority and Resources
5.2.4	Political Scrutiny
5.2.5	Ethical Concerns
5.2.6	Responsibility Attribution

Table 1. Categorization of the results

5.1 AI Implementation Processes in Dutch Public Sector Agencies

Outlining the implementation processes of Public Sector AI projects, provides the necessary information to contextualize the identified implementation challenges (section 5.2). We organize the AI implementation processes in phases based on the AI deployment life-cycle [13, 46], namely: *initiation*, *design- and development*, *piloting*, *deployment*, and *evaluation*[18]. The studied public sector agencies have dedicated innovation teams or labs to introduce data-driven techniques to the organization. Note that a large part of the AI Implementation process, occurs within- and with these teams. The respondents linked to those teams, agreed that the demonstration of AI prowess, can put AI on the map.

(1) *Initiation phase.* Initiating AI projects to include algorithms into decision-making processes, does not happen in a unified manner; the most heard ways are managers introducing them from strategic aspiration, or data scientists (often in cooperation with either managers or decision-makers) who started experimenting with AI opportunities (M1; M3; M5; DS3; DM2). Such experimentation either stems from questions to optimize business practices (e.g., detecting ships that get rid of chemical toxins on the coasts of Asian countries [19]), or starts from the possibilities of an algorithm, leaving questions for if and how the algorithms will be implemented for later: *"So the moment we had a few nice models, they started thinking; how do we get these integrated in work practices?"* (DS3). Several respondents mentioned desirable preconditions, to initiate the AI projects: practical considerations (time, priority), data-oriented considerations (sufficiency,

[17] Note that in the case of the ILT, the respondents refer to decision-makers as 'inspectors'.

[18] Even though the AI life-cycle can be defined in multiple ways, and in practice these phases are not linear, the studied cases could identify with these phases, therefore, these served to position the challenges, surfacing at different phases in the process.

[19] https://www.bnnvara.nl/zembla/artikelen/het-gifschip-van-sbm (last accessed 02.07.2024).

understanding) and ethical considerations (bias, impact) (M3). However, respondents also mentioned lack of some of these preconditions, feeding into challenges such as having a shared vision on the concept or the purpose of using AI, and difficulties to integrate AI into the current workflows, mentioned in section 2.2.

(2) *Design- and development phase.* The AI model's design choices and development, are still experimental, depending on the purpose of the model, feasibility for the data scientist, feasibility within the organization and motivation (M4; OB1 ; OB3). The complexity of models vary ("*So we use more complicated models such as random forests, but only in a use-case where explanations are very simple*" DS2). Regression models, decision trees and random forests, are mainly mentioned within the studied context. Note that the specifications of the models are known and mentioned by data scientists, and by two of the managers spoken to. They are not known by all managers, and by none of the decision-makers, introducing challenges such as difficulties to estimate the model's impact, and concerns about explainability.

(3) *Piloting phase.* At the time of research, two or three projects in each studied organization reached their pilot phase. Elucidating the often lengthy and uncertain process behind going from pilot to deployment: *"So we know we have good models, we (...) tested them, we (...) met one-to-one with inspectors, and (...) started the pilot. However, we haven't had enough feedback from the inspectors to say that the pilot has been (...) concluded. (...) If that will happen or not, that's far beyond me. (...) But I think it's also just the level we are right now"* (DS2) [20]. A combination of challenges mentioned in the next section (e.g., scarcity of decision-makers' input, political scrutiny, ethical concerns, unclear responsibility attribution), were given as reasons to hinder deployment.

(4) *Evaluation phase.* Evaluation of AI is done within the development teams, generally further along in the process; after development and often during or after the pilot phase. One of the studied cases increasingly includes a third-party in their evaluation (DS7). A decision-maker added the wish for a practitioner's check: *"Well, after the development we need to regard the following: Is the system really a lot better at filtering than a human being?"* (DM2). Evaluating the AI projects at the end of the process, and mostly within the development teams, cause that some of the challenges surface in a later stadium than they commence (e.g., resistance to change, concerns about fairness).

5.2 Multi-Actor Challenges during AI Implementation

Where some challenges were expressed by all actors (e.g., lack of priority and resources), others were perceived differently by the different actors (e.g., lack of shared understanding, responsibility attribution). Starting from different views on the concept and meaning of AI, some challenges are related to the key actors' varying roles and ways of perceiving — and interacting with — the systems. We highlight *Lack of Shared Understanding, Difficulties to Integrate AI in Current Workflows, Lack of Priority and Resources, Political Scrutiny, Ethical Concerns,* and *Unclear Responsibility Attribution* as challenges from a multi-actor perspective below (see Table 1).

5.2.1 *Lack of shared understanding among key actors.*

"What is AI?". Views around the meaning of AI, differed among respondents; ranging from regarding AI as a broad concept, to specific and more complex models - where Excel sheets are regarded 'old fashioned' or disregarded - to limited understanding of AI at all. The limited understanding of AI, perhaps notably, is only mentioned by managers about other managers, and

[20] Since the AI systems discussed with the respondents had not yet been deployed at the moment of research, we skip the deployment phase here, and proceed to the evaluation phase.

by decision-makers about themselves. Exemplifying this sentiment a decision-maker says: *“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* (DM3). Similarly, other decision-makers are unsure what AI means or how it works exactly (DM4; OB2). Forms of education are mentioned, to create more AI awareness and knowledge (M3).

The answer to “What is AI?” was not only different in *what* respondents answered, but also in *how* they answered. Data scientists and managers reasoned from the concept of AI or its context, then turning to AI in their- and the organization’s practices. The decision-makers on the contrary, reasoned from their practical workflows, then exploring if and how AI could fit there within. The discussions about AI with decision-makers were never unaccompanied by examples from practice (DM1; DM2; DM3; DM4). They often started with, perhaps seemingly unrelated, examples, but always working to a point where AI was compared to previously novel innovations that had been introduced to them. In other words, even if the definition was not clear to them, decision-makers could reason about the topic of AI, they would just approach the topic from a larger, more practical narrative.

"Why Should we use AI?". *Why* specific projects are initiated or used, is not always clear to respondents, or the answer is not necessarily shared by the different parties involved. *"Efficiency is mostly the main reason"* (M1), alleviating repetitive tasks or pre-selecting information are often mentioned as theoretic motivations to start AI projects. Likewise, the hope to improve quality and accuracy, were mentioned *“We want to move from inspections where they find nothing, to inspections where they find something. So (...) it is about efficiency, but mostly about quality”* (M3). That said, the value of using the AI system in practice is not always recognized. A data scientist explains: “*The project started with a question from the inspectors. (...) They were happy with the initial version but then they said: ‘okay, nice, we will see if we can use it*” (DS3). The decision-makers had different expectations about using the AI model than the data scientist involved.

"How should we use AI?". Considering that it is often unclear what AI may provide in the future, both managers and data scientists considered it useful to experiment with AI projects and create examples for the organization, at the same time learning if the models can perform well (M1; M2; M3; M5; DM1; DM2; DM3; DM7). *“The idea of our lab is also that we do experiments and show what is possible with data science, with AI”* (DS3). However, many questions about *how* to then develop, use and implement the AI systems, are up for debate(OB1; OB3). Some respondents talked about introducing AI in the current workflows (e.g. by having AI replace certain tasks done by humans). Others talked about introducing entirely new organizational workflows. Both streams of thought left many unanswered questions. How to make sure the algorithms are ethically sound? When should developers and other actors interact? What is the ideal interaction frequency? How to make sure the rest of the organizations uses the model, not only the people involved in their development? Taking an example where data scientists a-priori aimed to discuss modeling choices with decision-makers; the data scientists experienced a dilemma. It did not resonate with non-technical experts to discuss a still non-functioning model, then, how could they include the decision-makers’ input early and optimally in the process? (DS2).

Where every respondent mentioned that *the machine will not replace the human* (DS4), some thought a model’s decision should only be questioned if the result felt odd: *“If the computer says yes, then it is a yes. If it should have been a no in hind-sight, there is a mistake in the algorithm and you need to tweak it”* (M1). Others, mainly decision-makers, thought of the outputs of a model as a starting point, to be left entirely to human discretion to act upon (DM4). Others still, regarded the AI model as a *‘digital person’* (M2) who will closely interact with a *real person*, to decide if they accept the decision or not.

"What is the impact of AI?". Impact assessment is technically included in the form of review frameworks within one of our cases, in the other case it was not explicitly mentioned by the respondents. However, data scientists communicate that impact assessment is more than a technical matter. *"The model can have impact that we do not always foresee. We try to include the impact, but it is a difficult question"* (DS6). At the same time, they argue that *"a team manager might not always understand the model well enough to estimate its impact."* (DS3).

Managers argue that the current European precedents of tracing the impact of AI models, are largely based on the inputs of their own organization or the Dutch Ministries, therefore there is limited reference as to which impacts could be considered normal versus dangerous.

5.2.2 Difficulties to Integrate AI in Current Workflows.

Scarcity of Decision-Maker Input. Every data scientist (DS1-DS7) mentioned the difficulty in reaching (potential) users, i.e., the decision-makers. The importance of this, to improve the model's alignment with practice as well as increase the likeliness of adoption, is acknowledged. However, to structurally reach the decision-makers, proved difficult. Different reasons were mentioned, for instance: (1) Wanting to see if the model works before presenting it to the organization (mentioned by data scientists), (2) Limited time to do one's own work, let alone to add meetings and learning to work with new AI methods (mentioned by both data scientists and decision-makers alike), (3) Limited priority from managers to allocate this time (mentioned by data scientists and decision-makers), (4) Limited sense of trust, understanding or knowledge about the subject or urgency (mentioned by data scientists), (5) Conversations are much targeted to-, and spoken in the jargon of the data scientists, therefore they are experienced to be less useful for the decision-maker *"there is just really a mismatch between the models and the practice"* (DM3) (mentioned by decision-makers).

The differences in 'languages' between the different actors, formed a much expressed challenge for the AI implementation process in the cases. *"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"* (DS2), *"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"* (DS7). Moreover, being in physically different locations, was mentioned to enlarge the mental distance.

Interdependent Workflows. *"You need everything and everyone, and that makes this game very interesting but also frustrating at times"* (M4). Note that the studied Dutch public sector agencies exist in a complex and interrelated playing field with many different other agencies, ministries, government bodies, commercial parties and citizens. Inter-dependency then, refers to the dependencies between the actors already mentioned (e.g., managers are dependent on other managers, data scientists on decision-makers, decision-makers on managers, dependencies on the IT department). But it also refers to the multitude of public sector partners ('*ketenpartners*')[21] and other third party collaborations. *"You have a stream of collaborations with other organizations who do similar investigations"* (M1). Related to AI and data specifically, the different agencies use different formats, use different algorithms, and are not always informed about one another's innovations. As a result, dependencies on data, on third party information systems used, but also simply dependencies on information from other organizations whilst doing your job in the field, are mentioned (OB2; DE2; DE4).

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

Resistance to novel ways of working. Introducing AI and learning to work with it, requires adjustments beyond the algorithm. Those novel ways of working and their challenges become most apparent in the pilot phase. Since there is no clear alignment on what AI is, why we should use it, how AI should be incorporated and which workflows should be adapted or not (see section 5.2.1)), not everyone can adapt to these changes easily. For instance, one decision-maker valued the potential of using more predictive models, to move towards a *"more preventive way of working"* (DM3), with the rationale that *"preventing is better than curing"* (DM3). However, they said implementing such models probably triggers resistance to change with other decision-makers, since *"the decision-makers are currently not used to working that way"* (DM3). Another decision-maker voiced hesitance to incorporate AI indeed: *"Yes, 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? (...) 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"* (DM4). Several managers mentioned that there may be a loss of autonomy, or perceived autonomy by decision-makers, in the changes introduced by AI (M1, M4). *"Suddenly, a decision-maker who used to independently make a decision, now gets told by a system that something else may be smarter to do"* (M4). A decision-maker added to this point, saying: *"I think it is 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"* (DM2).

Note that two decision-makers pointed out that changing the current workflows, also means running into challenges not necessarily considered by data scientists or managers. For instance, since many public sector workflows revolve around laws and regulations, the decision-maker says in practice, *"we do not always know if we have the jurisdiction to act in this novel fashion"* (DM3).

5.2.3 Lack of Priority and Resources. The challenge of having insufficient resources, such as time, money, became apparent throughout the interviews, mostly when talking about the initiation of AI systems (section 5.1). The resource of 'time' was mentioned by almost every respondent interviewed, often in the context of the time-consuming process to include decision-makers (mentioned by data-scientists) or the lack of time decision-makers have to do their own jobs, let alone include new tasks (such as giving input for an AI project, mentioned by managers and decision-makers). One of the decision-makers explained that *"I recognize that we do not feel included on the one hand, but we also do not have time for it on the other hand"* (DM4).

Respondents pointed to a lack of priority; for managers to think about the implications of AI systems, for data scientists to incorporate organizational aspects beyond developing algorithms, and for decision-makers to contribute to the systems. Being involved in a new AI project *"should be doable if it is supported organization-wide. (...) But right now, I have my hands full, so it requires choices. (...) if the room is not made, I am also not going to do it."* (DM4). Team managers pointed to higher management or politics, to make strategic choices and allocate more time, resources or explicit responsibilities for such tasks. However, they identify a paradox, where: the value or pitfalls of the AI system are often demonstrated only after development and piloting, yet the resources allocated are often not sufficient to demonstrate this value in the first place.

5.2.4 Political Scrutiny. One major issue is the reluctance to implement AI due to the politically sensitive topics dealt with (see chapters 1 and 3), in turn increasing scrutiny as media may jump on it. A few statements to illustrate this: *"It would have been different if the organization was not so much scrutinized. (...) if the Deputy Minister says: I need to personally have an opinion about the topic, then that's the way it is"* (M1); *"Yes, often the current administration is decisive for what gets done at the ministry (...) bottom-up, we are not capable of taking out the political"* (DM4). The timing of interviewing proved of importance, as contextual issues put the organizations under close observation. *"So yes, you find yourself under a magnifying glass and, where the industry will* *get away with it, we will not"* (DS3). Issues like the *toeslagenaffaire* (the childcare benefit scandal already mentioned), a news article on the degassing of inland navigation ships (22nd January 2023) [22], and a Zembla episode (9th of September 2019)[23] on the ditching of chemical toxins on the coasts of Asian countries, all indicate increased societal and political pressure and observance (DM2; DS4). At the same time, respondents mentioned sub-surface political factors. One decision-maker said the fear of being reprimanded or being held accountable for an AI system you cannot completely oversee, makes people less inclined to participate (DM2). A data scientist talks about events of questioning of AI projects, in parliament or generally in the public debate: *"Either people think everything is dangerous, 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"* (DS7).

5.2.5 Ethical Concerns. Ethical concerns have been raised in various studies to increase responsible practices in AI, arguably more in the public sector context. Analogously relevant in the studied cases, respondents – managers and data scientists using similar terminology as in literature, decision-makers talking about concerns of a 'fair', 'understandable', or 'explainable' system intuitively – talked about ethical concerns to challenge the implementation of AI projects. The ethical concerns mentioned by respondents, considered topics of fairness and bias, data quality and explainability.

Concerns about Unfairness and Bias. The topic of fairness was discussed by all respondents in one case and by half of the respondents in the other *"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? (...) Defined by what we as a society deem sensitive"* (DS2). Most respondents considered the topic to be important, apart from two respondents, one manager and one decision-maker, indicating: *"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"* (DM2). That said, most respondents were concerned about unfairness, just not in the current stages of the development, arguing that initiatives have been put in place to estimate fairness implications later on. 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. Note that the different respondents used different words for- or mean different things by unfairness, ranging from more abstract or societal discussions to technical fairness methods. One respondent differed between bias and fairness, saying: *"When I say bias and ethics, these are more like high level; 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"* (DS2). For other, mostly non-technical, respondents there was less of a strict distinction and the concepts were often used interchangeably.

A field of tension can be detected here, between the challenges of unfairness and the difficulty of integrating AI into current workflows. Where many respondents – not necessarily of one actor group – vouched to align the AI model's input and output with the decision-makers' experiences to increase their adoption in current workflows, others saw an increased risk of attaining unfair outcomes. Respondents expressed the idea that having model outputs that align with the decision-makers' experiences, increases trust in the models, and therewith, increasing their willingness to adopt the AI systems (M2; DS1; DS2; DS3; DS4; DS5; DE2; DE3). Others however, argued that trusting a model alone, does not mean it is trustworthy. One of the data scientists said, *"Well if the score does not match their gut feeling, they will ignore it"* (DS6). That said, they critique to trust a

[22] https://www.nrc.nl/nieuws/2023/01/22/hoofdpijn-duizeligheid-en-tranende-ogen-dankzij-de-giftige-dampen-van-het-binnenvaartschip-a4154840 (last accessed 02.07.2024).

[23] https://www.bnnvara.nl/zembla/artikelen/het-gifschip-van-sbm (last accessed 02.07.2024).

model based on “gut-feeling”, as it leads to bias confirmation. A manager talked 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. But sometimes it is also right to have a level of distrust”* (M4).

Note that discussions are running about the most important features to include here and what the impact of these decisions would be (OB1; OB3). For instance, false positives and false negatives are both undesired outputs, yet with different outcomes. In one case, false positives would lead to investigating innocent decision subjects, whereas false negatives might have a negative impact on society; a trade-off the different respondents at the table valued differently.

Concerns about Data Quality: Sources, Scarcity and Historical Bias. Accurate and sufficient data is crucial for a responsibly working model. Problematically, however, such data is not always available. One of the data scientists mentioned this as one of their biggest challenges: *“For me, the biggest issue has to do with first, defining a problem to solve (...) and second, the data quality. There is no data. There is no big database that is structured where you can find things. (...) Often the data doesn’t even exist to make it robust”* (DS4) The core data challenges are to (1) get access to reliable data, especially due to the cooperation between- and dependency on other actors to share the data (as mentioned in section 5.2.2) (OB1; DM2). 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. (2) Make sure the data is sufficient, both for training and testing. Besides, COVID19 has caused a gap in available data and/or a distorted picture for the last couple of years. (3) Try to account for historical bias (OB1; OB3). Ideally, one randomly generates sampled data, to train and test the model. However, in practice, the data available is already influenced, based on years of decisions made by humans.

Concerns about the Effectiveness of Explainability. To give insight into the modeling choices and impact, one of the cases uses feature importance, as a tool to increase explainability. They also translated the feature importance into a dashboard for decision-makers. The explanations proved suitable to the language of data scientists, but less so for the decision-makers. A data scientist explained: *“We think that it is all very clear and useful for the decision-makers. For instance, all of us thought; a scatter plot, of course we should include that in our explanations, it is enlightening. But the decision-makers really did not have a clue. We [data scientists] think; it is so simple, even a correlation, we think it is the easiest thing there is, but (...) we need to think of something else to visualize or explain it”* (DS5). Moreover, the data scientist wondered if visualization is the only problem, or if it is the ML techniques themselves. *“It’s such a complex technique, can you really simplify it more? (...) to the extent that everyone can understand it. But it is the question then, if it still represents what lies underneath? (...) So, perhaps you need to use easier techniques. (...) make the choice between explainability or accuracy of machine learning, and perhaps go back to decision trees, or rule based, or something else simpler”* (DS5).

Two out of four decision-makers, were aware of the existence of the dashboard, but terms like ‘SHAP’[24] or ‘feature importance’ were not mentioned by them in the context of explainability. One decision-maker related effectiveness of explainability methods to direct verbal or written explanations from data scientists: *“well the explainability methods are sufficient now, in the pilot phase, since there is a close cooperation between a select group of interested data scientists and decision-makers. (...) However, once in production, the connections are less easy, and then explainability methods as used now, are not enough”* (DM3). Note that the question, *“If any explainability methods or type of*

[24] SHAP (SHapley Additive exPlanations) values are a well-know approach in XAI to measure the relative contribution of a given feature to the model’s output [59].

explainability are used for managers?" elicited an unanimous "no". Multiple respondents indicated that towards the manager, that type of detail level is not required (M1, M2, M3, DS2, DS3).

5.2.6 Unclear Responsibility Attribution. [25] Questions about responsibility: "1. Who should be responsible to make the call whether to go live with a pilot project?", "2. Who should have the responsibility to incorporate explainability, or other ethical values within a model?", and "3. Who should be responsible for deployment and maintenance?", were open-ended according to respondents (M2-M4; DS1-DS7; DM2; DM3). Taking an exemplifying answer to the first question: *"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 deploy. (...) you would want someone else to make the decision, but then really based on understanding the model and the facts, and not just a news article they read in the AD (Dutch newspaper), because that is what happens a lot now"* (DS7). The respondents reflected that the responsibilities are by no means clearly defined. For instance, is it the responsibility of the data scientist to include the perspective of the decision-maker in the development process? Some data scientists thought it is, whereas others thought of it beyond their job, or at least beyond their responsibility (DM2; DM3; DM6; DM7). Questions of ownership are raised. *"Right now, there is a big discussion about ownership in our organization and it's still undefined. I think (...) the final ownership will go to the end user. So the data scientist is responsible for the development. (...) Once the pilot is done, it goes to another team that supports the maintenance and they have ownership of the methods (...) 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"* (DS2).

6 DISCUSSION

In this section we summarize the key takeaways of our results and connect them to related literature. We identify implications for each of the discussed challenges, and highlight challenges and opportunities for future research.

6.1 Define Organizational Prerequisites before and during Initiation of an AI-system

Our findings suggest a discrepancy between the initiation- and development phases of an algorithmic system, and its adoption by the wider public organization. Throughout the AI implementation process, challenges emerged, some of which may be tackled if recognized and defined from the initiation phase onward. The AI models researched were considered less technologically complex, contrary to their governance and inter-actor implementation processes. As a result, rather than technical or model-specific challenges, socio-technical, political, and organizational challenges were emphasized by the key actors involved.

#Lack of Shared Understanding: Promote Synergy between AI Literacy and Domain Literacy. Many of our non-technical participants, namely 75% decision-makers and 50% of the managers, expressed to have no clear understanding of AI. Besides, there was a lack of clarity why or how AI should be used, i.e., there was no shared understanding of the goals and, thus, if they were achieved. Having these different understandings means there is a different baseline through which the actors operate and communicate. In turn, it affected decision-makers' willingness to give input during the development of the systems and to embrace new ways of working. It also affected how most technically-oriented participants expressed the wish to cooperate with other actors, yet did not know how to, and how managers struggled to estimate the AI projects' impacts and implications.

The organization might need to take a step back and discuss the different takes on AI and how these would fit into the organization. Our participants' experiences align with claims of

[25] By Responsibility Attribution, we refer to the question of *who* is responsible *for* something [23].

limited knowledge about ML and AI in public administration [78], and gaps in (AI) education [73] among employees of the public sector. Our results highlight the need to reduce the knowledge gap [103] and promote AI literacy among managers and decision-makers, though tailored to their respective roles. At least addressing when AI is being used, and when it is responsible to use AI [57]. A significant challenge is the ability to explain AI in non-technical terms and make technicalities understandable and intuitive for non data scientists [57]. Strategies could include designing pedagogic materials [101], yet based on practical use cases and real-world examples that relate to practitioner's existing ways of working [25, 42, 54]. Such strategies should emphasize the sociotechnical nature of AI, without limiting pedagogic resources to computational approaches [63].

Less acknowledged in the debate about improving AI literacy [73, 78], is improving domain literacy – to be included in the organizations' approach to AI. It is imperative to improve our understanding of the decision-makers' preferred interaction with AI. Note that it is not only the less technologically 'advanced', aware or inclined people who need to become AI literate, but also vice-versa: the organizational expertise and practical implications need to be incorporated into the algorithm systems, for them to be worthy of adoption.

#Difficulties to integrate AI in Current Workflows: Invest in AI that adds Value to Decision-makers. Many of the challenges to integrate AI in the current workflows, were closely tied to improving the users' –in this case decision-makers– input and experience. Scarcity of decision-makers' input and the resistance to novel ways of working with AI being the prime examples thereof. This resonates with previous work pointing to the lack of expert input in AI systems used for public decision-making [2, 103]. Our results highlight decision-makers' concerns about the limitations of AI systems [107], and their difficulties or struggles to understand algorithmic predictions [36, 47].

Decision-makers should be given the resources to progressively adapt their workflows to AI and to purposefully make use of this technology [113]. This ties to addressing the sources of skepticism of decision-makers, e.g. providing reliable data, enhancing the structural embedding of AI projects in the wider organization, and improving the tools to engage in fruitful, timely conversations. For instance, our interviewees highlighted the scarcity of reliable data, making it challenging to ensure data sufficiency both for training and testing. As pointed by previous work [103], even though the importance of relevant data and reaching consensus with key stakeholders regarding data sources is theoretically well known [90], in reality lack of access to information and issues regarding data privacy and data dependency, deter the responsible implementation of public sector AI projects. For effective collaborations between actors, it is key that actors with organizational expertise can engage in fruitful and timely conversations with actors with expertise in AI. A promising future research path is further exploring and developing *leaky abstractions* (i.e., ad-hoc representations exposing low-level implementation details [95]). These abstractions could include greybox prototypes or knobs to tune model parameters to achieve desired algorithmic behavior [95].

#Lack of Priority and Resources: Promote the Deliberation of Value Prioritization in AI Projects. Lack of priority and resources compromises the thoughtful design and implementation of AI systems, and might ultimately be a source of potential harms. Even before initiating an AI implementation project, the mere need of an AI system should be questioned and reflected upon [6]. This might require balancing taboo trade-offs [93, 97], which previous work has suggested should be done by society as a whole [93]. Unpacking the different underlying values of newly introduced AI projects – e.g., sigma values such as efficiency, and theta values such as transparency that may be competing with each other (see section 2.2.1), and understanding which lamda values are needed to make the organization more adaptive –, will lead to a more deliberated start of algorithmic decision-making systems. If the need and benefits of an AI system are weighed in favor of initiating the systems, a few key-elements need to be addressed. Our results suggest to at least: (1) Substantiate the initiation

of algorithms, (2) Design for a system that includes- educates, and tailors to the different actors from the start (including defining their responsibilities, elaborated in section 6.2), (3) Allocate sufficient resources to investigate, develop and implement responsible algorithmic systems, and most fundamentally (4) Transparently prioritize public values and their trade-offs.

6.2 Restoring Discretionary Power through Value-driven Transparency, Contestability and the Right to Alteration

Where some challenges can be tackled by anticipating them earlier in the AI implementation process, participants also emphasized to account for unpredictable effects of AI projects and unexpected challenges. This implies accounting for feedback loops in the process, and designing checks and balances for AI-human interacting public sector decision-making processes.

#Political Scrutiny: Uphold Transparency. The political sensitivity of AI in the Dutch public sector, increased significantly in recent years and raised concerns around AI implementation. Participants were wary of the risks associated with such scrutiny, managers most notably referring to questions around public accountability, and decision-makers worried they might be held accountable for decisions they could not oversee. This relates to findings by Kawakami et al. [50], who studied how power relations influence choices around design, adoption and use of AI. In this case; how agency managers experience pressures from other powerful institutions (e.g., legal systems, and private companies). Whilst high-visibility and public exposure can have learning effects through increasing awareness and alerting involved stakeholders of existing risks [7], to allow and embrace such scrutiny, transparency is essential. Both public transparency and transparency to auditors in public organization are not only ethically important but also necessary for democratic self-government [56].

Algorithm registries can set initial standards for transparency. The Algorithm Register from the City of Amsterdam, which provides an overview of AI systems and algorithms used by the municipality, is one example [34]. However, process-based approaches such as this one do not consider whether a given system meets ethical standards[26]. On the other hand, outcome-based approaches, such as model cards [71], have been criticized for overlooking important development steps. Value-based approaches, which consider the full range of values in AI systems, have been suggested as a solution [15].

Transparency is important not only during the implementation process but also when undesired situations arise. For example, after the municipality of Rotterdam used a discriminatory algorithm to profile people in the context of social welfare fraud,[27] it provided independent investigators with extensive information, including the AI model, training data, and user manuals. This allowed for learning effects and increased understanding of the system's inner workings.

#Ethical Concerns: Algorithm-in-the-loop Configurations & the Right to Contest. Many of the participants highlighted ethical concerns, including issues with unfair outputs (oftentimes referred to as an afterthought, to be dealt with once the system had been developed), and scarcity and bias in data. The respondents were additionally asked to reflect on the use of explainability within the implementation process (see Appendix A). With the help of explanations, decision-makers wanted to compare the results of the system with their own experiences to decide whether to trust the system. Even if explanations were developed with the hope of enabling decision-makers to exercise discretion, data scientists raised issues of confirmation bias.

[26] In a commentary to Floridi [34]'s work, Cath and Jansen [17] state that AI governance-by-database can be decontextualizing and depoliticizing, which reinforces what they call "ethics theater").

[27] https://www.lighthousereports.com/investigation/suspicion-machines/

The decision-makers we interviewed suggested that currently introduced explainable AI methods fell short to properly evaluate the algorithms. Such claims resonate with scholars advocating for a paradigm shift from explainable AI to evaluative AI. Miller [69] argued that explainable AI has failed to meet the purpose for which it was conceived: help decision-makers consider the appropriateness of AI recommendations and follow them when necessary (i.e., appropriate reliance [108]). Miller [69] further argued that the assumption that decision-makers will engage with explanations is unfounded — a claim that our results confirm. Our results align with previous work pointing to the lack of usefulness of feature-based explanations to empower non-experts [10].

Feature-based explanations might not be enough for decision-makers to exercise meaningful human control [18]. To effectuate useful explanations, we advocate (1) to ensure algorithm-in-the-loop configurations (i.e., human-AI configurations where the final decision is made by a person rather than an algorithm [39]). , and (2) to ensure contestability in AI (i.e., systems that are responsive to human intervention throughout their entire life-cycle[4]), with the right to alter the system's design. Algorithm-in-the-loop systems could consist of algorithms that present the arguments for- and against a decision [68], though a person would still make the final decision. When it comes to contestability, all actors, but decision-makers most prominently, need to be given resources to understand their role within the human-AI configuration [57], to critically interpret the output of the system and to actively influence or contest its outcome when needed [51]. Contestable AI is still a growing field (e.g., [4, 60, 89, 102, 114, 115]), and typically focuses on decision-subjects [61, 102, 115], rather than other actors in the AI implementation processes [3]. This offers an opportunity for future research to design interactions that allow for meaningful feedback loops. AI suggestions that are contested by decision-makers need to lead to serious deliberation, with the option of altering the system if need be. Securing that contestability is taken seriously, will require extending both legal as well as institutional means.

#Unclear Responsibility Attribution: Restoring Discretionary Power. Concerns were expressed about the lack of clear responsibility attribution and the unease (especially among decision-makers) about potentially being held accountable for the failures of a system they cannot effectively oversee. Currently, even though the organizations recognized the potential of explainability to ensure that decision-makers can exercise discretion, the methods are either technologically oriented, not developed at all, or visualized in non-intuitive way for decision-makers. There is, therefore, a need to come up with strategies that address the displacement of discretionary power [117]. Since data scientists are neither aware of nor trained for exercising such discretionary power through code [117], public agencies should be required to open design processes of AI technologies to deliberation. Design choices in AI systems come with policy implications, which require processes and spaces for expert decision-makers' input, public participation and political accountability [76]. To this end, third-party developers should disclose e.g., the objective that the system is optimized for, the way in which data is treated and leads to a determination, the assumptions behind the model choice or the interfaces that mediate the interaction of decision-makers with the AI system [76].

6.3 Limitations

There are a few methodological limitations of this study. (1) Even though the case studies were carefully selected, we do not claim to generalize over the entire public sector. Expanding the research to include other types of public bodies, would provide the ability to broaden the insights. (2) The limited availability of decision-makers due to their central operational roles in the analyzed case studies became apparent throughout the research. Our limited success in recruiting decision-makers can also be attributed to their self-expressed lack of AI knowledge, and their limited interest in AI-based solutions to support their work. (3) The managers interviewed differed in tasks and hierarchical levels. This had the advantage of creating better insight into the AI system's creation as a whole, but forms a limitation to compare the managers' perspectives. The data scientists as well have different tasks and specialties, but the hierarchical levels tend to be more equally distributed. (4) To get insight into the entire AI implementation workflow, future research would include the decision subjects as important actors that were out of scope now. (5) Categorizing people in a group itself, reduces the ability to incorporate their individual tasks, hierarchy and perspectives. (6) Conducting the interviews in the Dutch language (for the majority of interviews) was a choice, benefit and limitation. The choice was to speak to people in their native language where possible. The benefit was that people express themselves more freely, and since daily practices are also in Dutch, it is easier to communicate about them. The limitation is that translation means a certain degree of interpretation. The choice to translate quotes instead of keep them in the original language, was to avoid traceability to the respondents whose interview was conducted in English. (7) Due to the political context of the topics discussed, required a constant balance: aiming to report a level of detail to convey the current practices unfolding in the public sector, whilst at the same time adhering to the ethical guidelines and securing anonymity for all the parties involved.

7 CONCLUSION

In this paper, we reviewed the challenges of implementing AI in the public sector from a multi-actor perspective; interviewing civil servants within the roles of managers, data scientists, and decision-makers. To this end, we researched two national executive agencies, prominent in the Dutch public sector landscape. Though many organizations are increasingly technically adept, their governing and social maturity levels concerning AI vary. As a result, we found that the key-challenges, rather than programming and verifying complex models, were governance, design, and people related.

We identified that the studied agencies recognize potential for AI. However, there is a lack of clarity among the key actors on why or how AI should be used, and how to identify the impact of AI projects. The actors experience difficulties in adopting AI within the existing organizations' workflows. They identify a lack of priority, and insufficient resources. A first step to tackle these challenges, is to define clear organizational prerequisites before starting novel AI-related projects – e.g., allowing for non-technical actors to meaningfully contribute to the projects, and engaging in deliberating which potentially taboo-values to prioritize within AI projects.

Heightened by the politically delicate context, especially following prior negative experiences with the large-scale use of AI in the Dutch public sector, political scrutiny is referred to as a challenge to implement AI projects. Ethical concerns are raised, though the different actors highlight different concerns and ways to deal with them. All actors signal a lack of responsibility attribution, where their roles and responsibilities within AI projects are not clearly defined. We suggest to uphold transparency and therewith increase public accountability of the implementations. Moreover, we vouch to provide explanations especially catered to non-technical audiences, allowing to contest decisions throughout the implementation life-cycle, including altering or halting design-choices if necessary. Finally, establishing clear criteria for responsibility attribution will help to divide roles and responsibilities; permitting more targeted tackling of ethical concerns and retained discretionary power for decision-makers. As such, we do not only acknowledge the importance of considering ethical algorithmic values in the design of AI-systems, but offer a step forward. Towards the responsible adoption of algorithms, in societal contexts where the stakes are high.

REFERENCES

[1] Pankaj K. Agarwal. 2018. Public administration challenges in the world of AI and bots. *Public Administration Review* 78 (6 2018), 917–921. https://doi.org/10.1111/puar.12979

[2] Omar Saeed Al-Mushayt. 2019. Automating E-Government Services With Artificial Intelligence. *IEEE Access* 7 (2019), 146821–146829. https://doi.org/10.1109/ACCESS.2019.2946204

[3] Kars Alfrink, Ianus Keller, Neelke Doorn, and Gerd Kortuem. 2023. Contestable Camera Cars: A Speculative Design Exploration of Public AI That Is Open and Responsive to Dispute. In *Proceedings of the 2023 CHI conference on human factors in computing systems*. https://doi.org/10.1145/3544548.3580984

[4] Kars Alfrink, Ianus Keller, Gerd Kortuem, and Neelke Doorn. 2022. Contestable AI by Design: Towards a Framework. *Minds and Machines* (8 2022). https://doi.org/10.1007/s11023-022-09611-z

[5] Algemene Bestuursdienst. 2021. Dilemma's politiek en uitvoering; 'Alle lichten staan op groen. https://magazines.algemenebestuursdienst.nl/abdblad/2021/01/alle-lichten-staan-op-groen

[6] Ali Alkhatib and Michael Bernstein. 2019. Street-level algorithms: A theory at the gaps between policy and decisions. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–13. https://doi.org/10.1145/3290605.3300760

[7] Saar Alon-Barkat and Madalina Busuioc. 2023. Human–AI Interactions in Public Sector Decision Making: "Automation Bias" and "Selective Adherence" to Algorithmic Advice. *Journal of Public Administration Research and Theory* 33, 1 (2023), 153–169. https://doi.org/10.1093/jopart/muac007

[8] Amnesty International. 2021. Xenofobic Machines: Discrimination through unregulated use of algorithms in the Dutch childcare benefits scandal. https://www.amnesty.org/en/documents/eur35/4686/2021/en/

[9] Agathe Balayn, Mireia Yurrita, Jie Yang, and Ujwal Gadiraju. 2023. "Fairness Toolkits, A Checkbox Culture?" On the Factors That Fragment Developer Practices in Handling Algorithmic Harms. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society* (Montréal, QC, Canada) *(AIES '23)*. Association for Computing Machinery, New York, NY, USA, 482–495. https://doi.org/10.1145/3600211.3604674

[10] Astrid Bertrand, James R. Eagan, and Winston Maxwell. 2023. Questioning the ability of feature-based explanations to empower non-experts in robo-advised financial decision-making. In *2023 ACM Conference on Fairness, Accountability, and Transparency*. ACM, New York, NY, USA, 943–958. https://doi.org/10.1145/3593013.3594053

[11] Sarah Brayne. 2017. Big data surveillance: The case of policing. *American sociological review* 82, 5 (2017), 977–1008. https://doi.org/10.1177/0003122417725865

[12] Anna Brown, Alexandra Chouldechova, Emily Putnam-Hornstein, Andrew Tobin, and Rhema Vaithianathan. 2019. Toward Algorithmic Accountability in Public Services. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 1–12. https://doi.org/10.1145/3290605.3300271

[13] Jacob T. Browne, Saskia Bakker, Bin Yu, Peter Lloyd, and Somaya Ben Allouch. 2022. Trust in Clinical AI: Expanding the Unit of Analysis. *Frontiers in Artificial Intelligence and Applications* 354 (9 2022), 96–113. https://doi.org/10.3233/FAIA220192

[14] Lars Brummel, Sjors Overman, and Thomas Schillemans. 2021. Uitvoeringsorganisaties tussen staat en straat: De relevantie van maatschappelijke verantwoording voor directeuren van ZBO's en agentschappen. *Bestuurswetenschappen* 71 (1 2021), 27–45. https://doi.org/10.5553/Bw/016571942021075001003

[15] Stefan Buijsman. 2024. Transparency for AI systems: a value-based approach. *Ethics and Information Technology* 26, 2 (2024), 34.

[16] Justin Bullock, Matthew Young, and Yi-Fan Wang. 2020. Artificial intelligence, bureaucratic form, and discretion in public service. *Information Polity* 25, 4 (2020), 491–506. https://doi.org/10.3233/IP-200223

[17] Corinne Cath and Fieke Jansen. 2022. Dutch Comfort: The limits of AI governance through municipal registers. (2022).

[18] Luciano Cavalcante Siebert, Maria Luce Lupetti, Evgeni Aizenberg, Niek Beckers, Arkady Zgonnikov, Herman Veluwenkamp, David Abbink, Elisa Giaccardi, Geert-Jan Houben, Catholijn M. Jonker, Jeroen van der Hoven, Deborah Forster, and Reginald L. Lagendijk. 2023. Meaningful human control: actionable properties for AI system development. *AI Ethics* 3 (2023), 241–255. https://doi.org/10.1007/s43681-022-00167-3

[19] China Electronics Standardization Institute. 2018. Original CSET Translation of "Artificial Intelligence Standardization White Paper". https://cset.georgetown.edu/research/artificial-intelligence-standardization-white-paper/

[20] Charles West Churchman. 1967. Free for All. *Management Science* 14, 4 (12 1967), B–141–B–146. https://doi.org/10.1287/MNSC.14.4.B141

[21] Victoria Clarke and Virginia Braun. 2013. *Successful qualitative research: A practical guide for beginners.* Sage publications ltd. 1–400 pages.

[22] Victoria Clarke and Virginia Braun. 2021. *Thematic analysis: a practical guide.* SAGE Publications Ltd.

[23] Mark Coeckelbergh. 2020. Artificial intelligence, responsibility attribution, and a relational justification of explainability. *Science and engineering ethics* 26, 4 (2020), 2051–2068.

[24] Fariborz Damanpour and Marguerite Schneider. 2006. Phases of the adoption of innovation in organizations: effects of environment, organization and top managers. *British journal of Management* 17, 3 (2006), 215–236. https://doi.org/10.1111/j.1467-8551.2006.00498.x

[25] Wesley Hanwen Deng, Manish Nagireddy, Michelle Seng Ah Lee, Jatinder Singh, Zhiwei Steven Wu, Kenneth Holstein, and Haiyi Zhu. 2022. Exploring How Machine Learning Practitioners (Try To) Use Fairness Toolkits. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. ACM, New York, NY, USA, 473–484. https://doi.org/10.1145/3531146.3533113

[26] Wesley Hanwen Deng, Nur Yildirim, Monica Chang, Motahhare Eslami, Kenneth Holstein, and Michael Madaio. 2023. Investigating Practices and Opportunities for Cross-functional Collaboration around AI Fairness in Industry Practice. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT '23)*. Association for Computing Machinery, New York, NY, USA, 705–716. https://doi.org/10.1145/3593013.3594037

[27] Roel Dobbe, Thomas Krendl Gilbert, and Yonatan Mintz. 2021. Hard Choices in Artificial Intelligence. *Artificial Intelligence* 300 (2021), 103555. https://doi.org/10.1016/j.artint.2021.103555

[28] Zeynep Engin and Philip Treleaven. 2019. Algorithmic Government: Automating Public Services and Supporting Civil Servants in using Data Science Technologies. *Comput. J.* 62 (3 2019), 448–460. https://doi.org/10.1093/comjnl/bxy082

[29] The Human Environment and Transport Inspectorate. 2023. About the ILT. https://english.ilent.nl/about-the-ilt.

[30] European Commission. 2019. Ethics guidelines for trustworthy AI. https://www.aepd.es/sites/default/files/2019-12/ai-ethics-guidelines.pdf

[31] European Commission. 2024. EU Artificial Intelligence Act. https://artificialintelligenceact.eu/

[32] Jessica Fjeld, Nele Achten, Hannah Hilligoss, Adam Nagy, and Madhulika Srikumar. 2020. Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-Based Approaches to Principles for AI. *SSRN Electronic Journal* (2020). https://doi.org/10.2139/ssrn.3518482

[33] Luciano Floridi. 2019. Establishing the rules for building trustworthy AI. *Nature Machine Intelligence* 1, 6 (5 2019), 261–262. https://doi.org/10.1038/s42256-019-0055-y

[34] Luciano Floridi. 2020. Artificial intelligence as a public service: Learning from Amsterdam and Helsinki. *Philosophy & Technology* 33, 4 (2020), 541–546. https://doi.org/10.1007/s13347-020-00434-3

[35] Luciano Floridi, Josh Cowls, Thomas C. King, and Mariarosaria Taddeo. 2019. How to Design AI for Social Good: Seven Essential Factors. *Science and Engineering Ethics* 26 (2019), 1771–1796. https://doi.org/10.1007/s11948-020-00213-5

[36] Asbjørn Ammitzbøll Flügge. 2021. Perspectives from Practice: Algorithmic Decision-Making in Public Employment Services. In *Companion Publication of the 2021 Conference on Computer Supported Cooperative Work and Social Computing*. ACM, New York, NY, USA, 253–255. https://doi.org/10.1145/3462204.3481787

[37] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2021. Datasheets for datasets. *Commun. ACM* 64, 12 (12 2021), 86–92. https://doi.org/10.1145/3458723

[38] Google. 2018. AI at Google: Our Principles. https://www.blog.google/technology/ai/ai-principles/

[39] Ben Green and Yiling Chen. 2019. The Principles and Limits of Algorithm-in-the-Loop Decision Making. In *Proceedings of the ACM on Human-Computer Interaction*, Vol. 3. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3359152

[40] Stephan Grimmelikhuijsen and Albert Meijer. 2022. Legitimacy of algorithmic decision-making: six threats and the need for a calibrated institutional response. *Perspectives on Public Management and Governance* 5, 3 (2022), 232–242. https://doi.org/10.1093/ppmgov/gvac008

[41] Amy Heger, Elizabeth B. Marquis, Mihaela Vorvoreanu, Hanna Wallach, and Jennifer Wortman Vaughan. 2022. Understanding Machine Learning Practitioners' Data Documentation Perceptions, Needs, Challenges, and Desiderata. In *Proceedings of the ACM on Human-Computer Interaction*, Vol. 6. https://doi.org/10.1145/3555760

[42] Kenneth Holstein, Jennifer Wortman Vaughan, Hal Daumé, Miro Dudik, and Hanna Wallach. 2019. Improving Fairness in Machine Learning Systems: WhatDoIndustry Practitioners Need?. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 1–16. https://doi.org/10.1145/3290605.3300830

[43] Naja Holten Møller, Irina Shklovski, and Thomas T. Hildebrandt. 2020. Shifting Concepts of Value: Designing Algorithmic Decision-Support Systems for Public Services. *Association for Computing Machinery*, Article 70 (2020), 12 pages. https://doi.org/10.1145/3419249.3420149

[44] Christopher Hood. 1991. A public management for all seasons? *Public administration* 69 (1 1991), 3–19. https://doi.org/10.1111/j.1467-9299.1991.tb00779.x

[45] IBM. 2019. IBM Everyday Ethics for AI. https://www.ibm.com/watson/assets/duo/pdf/everydayethics.pdf

[46] ICO. 2023. Annex A: Fairness in the AI lifecycle . *ICO* (2023). https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/artificial-intelligence/guidance-on-ai-and-data-protection/annex-a-fairness-in-the-ai-lifecycle/?q=Auditing+Framework

[47] Marijn Janssen, Martijn Hartog, Ricardo Matheus, Aaron Yi Ding, and George Kuk. 2020. Will Algorithms Blind People? The Effect of Explainable AI and Decision-Makers' Experience on AI-supported Decision-Making in Government. *Social Science Computer Review* 40, 2 (4 2020), 478–493. https://doi.org/10.1177/0894439320980118

[48] Anna Jobin, Marcello Ienca, and Effy Vayena. 2019. The global landscape of AI ethics guidelines. *Nature Machine Intelligence* 1, 9 (9 2019), 389–399. https://doi.org/10.1038/s42256-019-0088-2

[49] Andreas Kaplan and Michael Haenlein. 2019. Siri, Siri, in my hand: Who's the fairest in the land? On the interpretations, illustrations, and implications of artificial intelligence. *Business horizons* 62, 1 (2019), 15–25. https://doi.org/10.1016/j.bushor.2018.08.004

[50] Anna Kawakami, Amanda Coston, Hoda Heidari, Kenneth Holstein, and Haiyi Zhu. 2024. Studying Up Public Sector AI: How Networks of Power Relations Shape Agency Decisions Around AI Design and Use. *Proceedings of the ACM on Human-Computer Interaction* CSCW2 (2024), 1–24. https://doi.org/10.1145/3686989

[51] Daniel N Kluttz, Nitin Kohli, and Deirdre K Mulligan. 2022. Shaping our tools: Contestability as a means to promote responsible algorithmic decision making in the professions. In *Ethics of Data and Analytics*. Auerbach Publications, 420–428.

[52] Daan Kolkman. 2020. The usefulness of algorithmic models in policy making. *Government Information Quarterly* 37, 3 (2020), 101488. https://doi.org/10.1016/j.giq.2020.101488

[53] Maciej Kuziemski and Gianluca Misuraca. 2020. AI governance in the public sector: Three tales from the frontiers of automated decision-making in democratic settings. *Telecommunications Policy* 44 (July 2020). Issue 6. https://doi.org/10.1016/j.telpol.2020.101976

[54] Michelle Seng Ah Lee and Jat Singh. 2021. The Landscape and Gaps in Open Source Fairness Toolkits. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 1–13. https://doi.org/10.1145/3411764.3445261

[55] Michael Lipsky. 1980. *Street Level Bureaucracy: Dilemmas of the Individual in Public Services*. Russell Sage Foundation.

[56] Michele Loi and Matthias Spielkamp. 2021. Towards accountability in the use of artificial intelligence for public administrations. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. 757–766.

[57] Duri Long and Brian Magerko. 2020. What is AI Literacy? Competencies and Design Considerations. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 1–16. https://doi.org/10.1145/3313831.3376727

[58] Kim Loyens and Jeroen Maesschalck. 2010. Toward a theoretical framework for ethical decision making of street-level bureaucracy: Existing models reconsidered. *Administration & Society* 42, 1 (2010), 66–100. https://doi.org/10.1177/0095399710362524

[59] Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. *Advances in neural information processing systems* 30 (2017). https://doi.org/10.48550/arXiv.1705.07874

[60] Henrietta Lyons, Eduardo Velloso, and Tim Miller. 2021. Conceptualising Contestability: Perspectives on Contesting Algorithmic Decisions. In *Proceedings of the ACM on Human-Computer Interaction*, Vol. 5. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3449180

[61] Henrietta Lyons, Senuri Wijenayake, Tim Miller, and Eduardo Velloso. 2022. What's the Appeal? Perceptions of Review Processes for Algorithmic Decisions. In *CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 1–15. https://doi.org/10.1145/3491102.3517606

[62] Michael Madaio, Lisa Egede, Hariharan Subramonyam, Jennifer Wortman Vaughan, and Hanna Wallach. 2022. Assessing the Fairness of AI Systems: AI Practitioners' Processes, Challenges, and Needs for Support. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW1 (3 2022), 1–26. https://doi.org/10.1145/3512899

[63] Michael Madaio, Shivani Kapania, Rida Qadri, Ding Wang, Andrew Zaldivar, Remi Denton, and Lauren Wilcox. 2024. Learning about Responsible AI On-The-Job: Learning Pathways, Orientations, and Aspirations. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*. 1544–1558. https://doi.org/10.1145/3630106.3658988

[64] Rohit Madan and Mona Ashok. 2023. AI adoption and diffusion in public administration: A systematic literature review and future research agenda. *Government Information Quarterly* 40, 1 (1 2023), 101774. https://doi.org/10.1016/j.giq.2022.101774

[65] Albert Meijer, Lukas Lorenz, and Martijn Wessels. 2021. Algorithmization of bureaucratic organizations: Using a practice lens to study how context shapes predictive policing systems. *Public Administration Review* 81, 5 (2021), 837–846. https://doi.org/10.1111/puar.13391

[66] Albert Meijer and Martijn Wessels. 2019. Predictive Policing: Review of Benefits and Drawbacks. *International Journal of Public Administration* 42 (12 2019), 1031–1039. https://doi.org/10.1080/01900692.2019.1575664

[67] Microsoft. 2018. AI Principles. https://www.microsoft.com/en-us/ai/responsible-ai?activetab=pivot1%3aprimaryr6

[68] Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence* 267 (2 2019), 1–38. https://doi.org/10.1016/j.artint.2018.07.007

[69] Tim Miller. 2023. Explainable AI is Dead, Long Live Explainable AI! Hypothesis-Driven Decision Support Using Evaluative AI. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency* (Chicago, IL, USA) *(FAccT '23)*. Association for Computing Machinery, New York, NY, USA, 333–342. https://doi.org/10.1145/3593013.3594001

[70] Gianluca Misuraca, Colin van Noordt, and Anys Boukli. 2020. The use of AI in public services: Results from a preliminary mapping across the EU. In *Proceedings of the 13th international conference on theory and practice of*

electronic governance. 90–99. https://doi.org/10.1145/3428502.3428513

[71] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2018. Model Cards for Model Reporting. (10 2018). https://doi.org/10.1145/3287560.3287596

[72] Brent Mittelstadt. 2019. Principles alone cannot guarantee ethical AI. *Nature Machine Intelligence* 1, 11 (2019), 501–507. https://doi.org/10.1038/s42256-019-0114-4

[73] Laura Montoya and Pablo Rivas. 2019. Government AI Readiness Meta-Analysis for Latin America And The Caribbean. In *2019 IEEE International Symposium on Technology and Society (ISTAS)*. 1–8. https://doi.org/10.1109/ISTAS48451.2019.8937869

[74] Mark Moore. 1997. *Creating public value: Strategic management in government.* Harvard university press.

[75] Jessica Morley, Luciano Floridi, Libby Kinsey, and Anat Elhalal. 2020. From What to How: An Initial Review of Publicly Available AI Ethics Tools, Methods and Research to Translate Principles into Practices. *Science and Engineering Ethics* 26 (2020), 2141–2168. https://doi.org/10.1007/s11948-019-00165-5

[76] Deirdre K Mulligan and Kenneth A Bamberger. 2019. Procurement as policy: Administrative process for machine learning. *Berkeley Tech. LJ* 34 (2019), 773.

[77] OECD. 2019. Recommendation of the Council on Artificial Intelligence. https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0406

[78] Adegboyega Ojo, Sehl Mellouli, and Fatemeh Ahmadi Zeleti. 2019. A Realist Perspective on AI-Era Public Management*. In *Proceedings of the 20th Annual International Conference on Digital Government Research* (Dubai, United Arab Emirates) *(dg.o 2019)*. Association for Computing Machinery, New York, NY, USA, 159–170. https://doi.org/10.1145/3325112.3325261

[79] Arthur C Petersen, Albert Cath, Maria Hage, Eva Kunseler, and Jeroen P van der Sluijs. 2011. Post-normal science in practice at the Netherlands Environmental Assessment Agency. *Science, Technology, & Human Values* 36, 3 (2011), 362–388. https://doi.org/10.1177/016224391038579

[80] Adamantia Rachovitsa and Niclas Johann. 2022. The human rights implications of the use of AI in the digital welfare state: Lessons learned from the Dutch SyRI case. *Human Rights Law Review* 22 (2 2022), 1–15. https://doi.org/10.1093/hrlr/ngac010

[81] Inioluwa Deborah Raji, I Elizabeth Kumar, Aaron Horowitz, and Andrew D Selbst. 2022. The Fallacy of AI Functionality. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. 959–972. https://doi.org/10.1145/3531146.3533158

[82] Bogdana Rakova, Jingying Yang, Henriette Cramer, and Rumman Chowdhury. 2021. Where Responsible AI meets Reality. In *Proceedings of the ACM on Human-Computer Interaction*, Vol. 5. 1–23. https://doi.org/10.1145/3449081

[83] Rijksoverheid. 2019. Strategisch Actieplan voor Artificiële Intelligentie. https://open.overheid.nl/documenten/ronl-e14cdcee-690c-4995-9870-fa4141319d6f/pdf.

[84] Rijksoverheid. 2022. Staat van de Uitvoering. https://www.rijksoverheid.nl/documenten/kamerstukken/2023/01/18/bijlage-3-staat-van-de-uitvoering-2022

[85] Rijksoverheid. 2023. Over de rijksinspecties. https://www.rijksinspecties.nl/over-de-inspectieraad/over-de-rijksinspecties

[86] Jeanne S Ringel, Dana Schultz, Joshua Mendelsohn, Stephanie Brooks Holliday, Katharine Sieck, Ifeanyi Edochie, and Lauren Davis. 2018. Improving child welfare outcomes: balancing investments in prevention and treatment. *Rand health quarterly* 7, 4 (2018).

[87] Horst Rittel and Melvin Webber. 1973. Dilemmas in a General Theory of Planning. *Policy Sciences* (1973), 155—-169. https://doi.org/10.1007/BF01405730

[88] M Rogers Everett. 1995. Diffusion of innovations. *New York* 12 (1995).

[89] Claudio Sarra. 2020. Put Dialectics into the Machine: Protection against Automatic-decision-making through a Deeper Understanding of Contestability by Design. *Global Jurist* 20, 3 (10 2020). https://doi.org/10.1515/gj-2020-0003

[90] Devansh Saxena, Karla Badillo-Urquiola, Pamela J. Wisniewski, and Shion Guha. 2021. A Framework of High-Stakes Algorithmic Decision-Making for the Public Sector Developed through a Case Study of Child-Welfare. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (10 2021), 1–41. https://doi.org/10.1145/3476089

[91] Kristen M. Scott, Sonja Mei Wang, Milagros Miceli, Pieter Delobelle, Karolina Sztandar-Sztanderska, and Bettina Berendt. 2022. Algorithmic Tools in Public Employment Services: Towards a Jobseeker-Centric Perspective. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. ACM, New York, NY, USA, 2138–2148. https://doi.org/10.1145/3531146.3534631

[92] Cathrine Seidelin, Therese Moreau, Irina Shklovski, and Naja Holten Møller. 2022. Auditing Risk Prediction of Long-Term Unemployment. *Proceedings of the ACM on Human-Computer Interaction* 6, GROUP (1 2022), 1–12. https://doi.org/10.1145/3492827

[93] Friso Selten and Albert Meijer. 2021. Managing algorithms for public value. *International Journal of Public Administration in the Digital Age* 8 (1 2021), 1–16. https://doi.org/10.4018/IJPADA.20210101.oa9

[94] Catherine Stinson. 2022. Algorithms are not neutral. *AI ethics* 2 (2022), 763–770. https://doi.org/10.1007/s43681-022-00136-w

[95] Hariharan Subramonyam, Jane Im, Colleen Seifert, and Eytan Adar. 2022. Solving Separation-of-Concerns Problems in Collaborative Design of Human-AI Systems through Leaky Abstractions. In *CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 1–21. https://doi.org/10.1145/3491102.3517537

[96] Tara Qian Sun and Rony Medaglia. 2019. Mapping the challenges of Artificial Intelligence in the public sector: Evidence from public healthcare. *Government Information Quarterly* 36, 2 (4 2019), 368–383. https://doi.org/10.1016/J.GIQ.2018.09.008

[97] Philip E. Tetlock, Orie V. Kristel, S. Beth Elson, and Melanie C. Green. 2000. The psychology of the unthinkable: taboo trade-offs, forbidden base rates, and heretical counterfactuals. *Journal of personality and social psychology* 78, 5 (2000), 853–870. https://doi.org/10.1037/0022-3514.78.5.853

[98] Halil Toros and Daniel Flaming. 2018. Prioritizing homeless assistance using predictive algorithms: an evidence-based approach. *Cityscape* 20, 1 (2018), 117–146.

[99] Steven Umbrello and Ibo van de Poel. 2021. Mapping value sensitive design onto AI for social good principles. *AI and Ethics* 1, 3 (8 2021), 283–296. https://doi.org/10.1007/s43681-021-00038-3

[100] National Science United States Executive Office of the President and Technology Council Committee on Technology. 2016. Preparing for the Future of Artificial Intelligence. https://obamawhitehouse.archives.gov/sites/default/files/whitehouse_files/microsites/ostp/NSTC/preparing_for_the_future_of_ai.pdf

[101] Kristen Vaccaro, Christian Sandvig, and Karrie Karahalios. 2020. "At the End of the Day Facebook Does What ItWants". *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW2 (10 2020), 1–22. https://doi.org/10.1145/3415238

[102] Kristen Vaccaro, Ziang Xiao, Kevin Hamilton, and Karrie Karahalios. 2021. Contestability For Content Moderation. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (10 2021), 1–28. https://doi.org/10.1145/3476059

[103] David Valle-Cruz, Edgar Alejandro Ruvalcaba-Gomez, Sandoval-Almazan, Rodrigo, and J. Ignacio Criado. 2019. A Review of Artificial Intelligence in Government and its Potential from a Public Policy Perspective. In *Proceedings of the 20th annual international conference on digital government research*. 91–99. https://doi.org/10.1145/3325112.3325242

[104] Marvin Van Bekkum and Frederik Zuiderveen Borgesius. 2022. Digital welfare fraud detection and the Dutch SyRI judgment. *European Journal of Social Security* 23 (4 2022), 323–340. https://doi.org/10.1177/13882627211031257

[105] Colin van Noordt and Gianluca Misuraca. 2022. Artificial intelligence for the public sector: results of landscaping the use of AI in government across the European Union. *Government Information Quarterly* 39 (2022). Issue 3. https://doi.org/10.1016/j.giq.2022.101714

[106] Anne Fleur van Veenstra, Francisca Grommé, and Somayeh Djafari. 2021. The use of public sector data analytics in the Netherlands. *Transforming Government: People, Process and Policy* 15 (2021), 396–419. Issue 4. https://doi.org/10.1108/TG-09-2019-0095

[107] Michael Veale, Max Van Kleek, and Reuben Binns. 2018. Fairness and Accountability Design Needs for Algorithmic Support in High-Stakes Public Sector Decision-Making. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 1–14. https://doi.org/10.1145/3173574.3174014

[108] Oleksandra Vereschak, Gilles Bailly, and Baptiste Caramiaux. 2021. How to Evaluate Trust in AI-Assisted Decision Making? A Survey of Empirical Methodologies. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (10 2021), 1–39. https://doi.org/10.1145/3476068

[109] Guido Vonk and Stan Geertman. 2008. Improving the adoption and use of planning support systems in practice. *Applied Spatial Analysis and Policy* 1 (2008), 153–173. https://doi.org/10.1007/s12061-008-9011-7

[110] Haiko van der Voort, A.J Klievink, Michela Arnaboldi, and Albert Meijer. 2019. Rationality and politics of algorithms. Will the promise of big data survive the dynamics of public decision making? *Government Information Quarterly* 36 (1 2019), 27–38. https://doi.org/10.1016/j.giq.2018.10.011

[111] Yao Xie, Melody Chen, David Kao, Ge Gao, and Xiang Antony Chen. 2020. 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. 1–13. https://doi.org/10.1145/3313831.3376807

[112] D Yang, B Qu, and P Cudre-Mauroux. 2020. Location-Centric Social Media Analytics: Challenges and Opportunities for Smart Cities. *IEEE Intelligent Systems* (2020), 1. https://doi.org/10.1109/MIS.2020.3009438

[113] Qian Yang, Aaron Steinfeld, Carolyn Rosé, and John Zimmerman. 2020. Re-examining Whether, Why, and How Human-AI Interaction Is Uniquely Difficult to Design. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 1–13. https://doi.org/10.1145/3313831.3376301

[114] Mireia Yurrita, Agathe Balayn, and Ujwal Gadiraju. 2023. Generating Process-Centric Explanations to Enable Contestability in Algorithmic Decision-Making: Challenges and Opportunities. (5 2023). https://doi.org/10.48550/arXiv.2305.00739

[115] Mireia Yurrita, Tim Draws, Agathe Balayn, Dave Murray-Rust, Nava Tintarev, and Alessandro Bozzon. 2023. Disentangling Fairness Perceptions in Algorithmic Decision-Making: the Effects of Explanations, Human Oversight, and Contestability. In *2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*. ACM, New York, NY, USA. https://doi.org/10.1145/3544548.3581161

[116] Mireia Yurrita, Dave Murray-Rust, Agathe Balayn, and Alessandro Bozzon. 2022. Towards a multi-stakeholder value-based assessment framework for algorithmic systems. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. ACM, New York, NY, USA, 535–563. https://doi.org/10.1145/3531146.3533118

[117] Stavros Zouridis, Marlies van Eck, and Mark Bovens. 2020. *Automated Discretion.* Palgrave Macmillan, Cham, 313–329. https://doi.org/10.1007/978-3-030-19566-3_20

[118] Anneke Zuiderwijk, Yu-Che Chen, and Fadi Salem. 2021. Implications of the use of artificial intelligence in public governance: A systematic literature review and a research agenda. *Government Information Quarterly* 38, 3 (2021), 101577. https://doi.org/10.1016/j.giq.2021.101577

A APPENDIX: 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 challenges experienced in AI implementation
  - 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
  - Ideas to improve current AI implementation
  - The potential for explainability to tackle implementation challenges
  - Ideas to improve XAI