Sarah Sherif Fathalla

Ethical Implications of AI Use in the Public Sector: An Exploratory Research on Dutch Citizens' Perspectives

Master Thesis

at the Chair for Information Systems and Information Management
(Westfälische Wilhelms-Universität, Münster)

Supervisor: Dr. Bettina Distel

Presented by: Sarah Sherif Fathalla

Date of Submission: 2023-06-03

Content

Figures

Tables

Abbreviations

AI	Artificial Intelligence
DESI	Digital Economy and Society Index
EC	European Commission
EU	European Union
HLEG AI	High-Level Expert Group on Artificial Intelligence
PSO	Public Sector Organization
SLR	Systematic Literature Review
UK	United Kingdom

1 Introduction

Discourse surrounding the use and application of Artificial Intelligence (AI) has garnered increasing attention in the past few years (Boyd and Wilson 2017). Multiple competing narratives exist over its use and impact on society. For instance, a Forbes article titled *Artificial Intelligence for Good: How AI is Helping Humanity* argues that AI is one of the most revolutionary human developments to exist, and that it is capable of addressing complex societal challenges (Sukhadeve 2021). Conversely, in an article titled *Artificial Intelligence Will 'Likely' Destroy Humans, Researchers Say*, Kokkinidis (2022) highlights that using AI may endanger society as it poses an existential risk. Media discourse thus attempts to idealise or diminish what AI as a concept and its applications entail, rather than addressing its core. This creates ambiguity, including how AI is understood and perceived by the general public.

Research regarding AI emerged in the 1940s, with studies focusing on the possibilities of decision-making by machines (e.g., McCulloch and Pitts 1943). Following that, general areas where AI solutions can be applied were searched for in the 1970s (e.g., Waterman and Newell 1971). Nowadays, research has expanded its focus to analysing the opportunities and applications of AI in public sector functions (Pan 2016). Relevant domains include education (Chen et al. 2020), public health (Benke and Benke 2018) and national security (Radulov 2019).

AI can be defined as autonomous systems operating in the absence of human mediation, that learn through identifying patterns in data to make decisions and realise different conclusions based on the analyses of various contexts (Čerka et al. 2017). The self-learning algorithm, which is the foundation for AI, is key for the emergence of innovations within different sectors in society (Wirtz et al. 2019). A report by the McKinsey Global Institute reiterates this assumption by highlighting that "rapid advances in automation and artificial intelligence have a significant impact on the way we work and our productivity" (Batra et al. 2018, p. 4).

Such assumptions prompted investments in AI-based technologies within the public sectors of various countries worldwide (de Sousa et al. 2019). For instance, The People's Republic of China (2017) has committed to investing approximately $150 billion to establish itself as a frontier in AI by 2030. In the United Kingdom (UK), the implementation of AI is anticipated to contribute £200bn to the country's gross domestic product by 2030 and create 80,000 new jobs, highlighting the public sector's willingness to transform established practices (Mikhaylov et al. 2018).

The introduction of AI as a novel technology in the public sector is often done under the guise of providing benefits to end users, which within this sector includes citizens. As the influence of AI has transformative impacts on social, political, economic and ethical elements, the incorporation of citizens' perceptions on the use of AI in the public sector is considered important (Chen et al. 2021). Their perceptions are relevant as the public sector operates under the mandate of providing public value, alongside achieving economic objectives (Rosemann et al. 2021). Despite this, citizens find it difficult to form definite opinions as AI implementations are considered "invisible elements of daily life, mostly driven by proprietary algorithms" (Yigitcanlar et al. 2022, p. 2). Accordingly, various elements pertaining to AI are difficult to grasp, making it harder for citizens to shape their opinions accurately.

As such, government authorities call for increasing discussions on AI advances and implementations in public discourse to engage and educate citizens about its use (Lee et al. 2020). This attempts to overcome what Crawford and Calo (2016) refer to as the blind spot in AI research – the lack of citizen involvement. Understanding citizens' perceptions of AI use in the public sector is relevant, particularly when upholding values that underpin democratic societies. While the use of AI aims to achieve more efficient outcomes, Hildebrandt (2016) highlights that democratic elements such as pluralism and upholding conflicting societal views ought to be considered too. In doing so, the risk of rising democratic paternalism perpetrated by government officials when chasing efficiency benefits is diminished.

Additionally, social scientific research iterates the intertwined link between technological progression and social acceptance. As Zhai et al. (2020) stated: "public perceptions and concerns about AI are important because the success of any emergent technology depends in large on public acceptance" (p. 140). Citizens' perceptions thus impact the adoption and scale of AI, which has profound impacts on how we structure society (Bao et al. 2022). This provides public administrators with the motivation to incorporate citizens' perspectives within AI discussions in order to enhance the successful implementation of AI in the public sector.

Ingrams et al. (2022) argue that there are two conflicting dimensions at play which impact citizens' perceptions on the use of AI. On the one hand, the instrumental dimension highlights that AI use in public sector organizations (PSOs) delivers greater efficiency benefits as opposed to traditional established methods and strategies (Young et al. 2019). Citizens hold optimistic views regarding the opportunities that AI offers, particularly when focusing on benefits related to time-saving when performing administrative tasks, such as filing taxes (Starke and Lünich 2020). On the other hand, the value-based

dimension entails that citizens consider the social, normative, and political implications of AI use in the public sector. Such value-based judgements are heightened when citizens consider elements such as faceless decision-making done by AI, which may be perceived as illegitimate interference or power attributed to AI (Busuioc 2022). These elements may adversely impact citizens' perceptions of ill-outcomes resulting of AI use, including inequalities and consistent biases (Easton 2018).

Accordingly, based on the aforementioned instrumental and value-based dimensions, the public sector finds itself in a predicament – being tempted by the promised efficiency benefits from implementing AI while protecting citizens from potential negative consequences of its use (Kuziemski and Misuraca 2020). The latter is important to consider as AI programs are adopted over diverse application areas, despite these programs posing risks of social destruction. Social destruction entails negative societal and organizational consequences that are considered undesirable (Newell and Marabelli 2015). This is particularly salient within the public sector context, as impacted stakeholders include citizens who are unable to opt-out of interacting with organizations implementing AI.

For instance, the use of AI in immigration enforcement in the UK has led to the cancellation of thousands of visas for immigrants due to a system error (McDonald 2020). Other examples include negative consequences arising from using AI in application screenings of university exams (Hao 2020) and for recidivism prediction in justice systems (Buranyi 2017), which brought about controversy due to heightened racial and socio-economic discrimination. A Gartner (2018) report predicts that erroneous outcomes may emerge in 85 percent of AI projects as a result of biased data or incorrect management. As such, this calls into question the government's ability to ensure citizens' rights are safeguarded when deploying AI (Crawford and Calo 2016).

As a response to the recognition of AI's transformative impact on various societal domains, debates surrounding the need for principles that safeguard values emerged to guide the development and use of AI. Worldwide, an ever-increasing number of organizations developed ethical AI frameworks and principles that enforce more careful implementation of AI to address the emergence of negative consequences from its use (Kuziemski and Misuraca 2020). This is part of the larger responsible business agenda that prioritises good governance and upholds societal concerns. Chen et al. (2021) argue that these guidelines and principles are meant to serve as a tool to increase citizen trust in AI systems, which impact their experiences with AI programs.

1.1 Outline of the Problem

Jobin et al. (2019) emphasise that in recent years, the heightened proliferation of ethical AI frameworks and principles in the public sector is seen as propagating soft-law efforts to address advances in AI use and implementation. Accordingly, national and international organizations created committees specialised in AI, with a mandate to produce guidelines and policy reports. Such committees include the Select Committee on Artificial Intelligence of the UK House of Lords and the High-Level Expert Group on Artificial Intelligence (HLEG AI) of the European Commission (EC). Due to the increasing diversity of guidelines and principles published, Floridi et al. (2018) aimed to produce a comprehensive ethical AI framework encompassing the most cited ethical AI principles – the AI4People Framework.

This was developed following a synthesis of reports published by six trusted initiatives and organizations. The five resulting principles were common across the reports analysed: beneficence, non-maleficence, autonomy, justice and explicability. Ashok et al. (2022) further the framework by mapping the principles on an ontological framework which consists of three domains – physical, cognitive, information – and a fourth governance domain added by the authors. In addition, they detail various ethical implications pertaining to each of the ethical AI principles. The resultant framework – Extended AI4People Framework – thus shows which ontological domains the principles impact, alongside their more detailed ethical implications as shown by research.

Despite the perceived positive turn of developing ethical AI frameworks, they are not without critique. Scrutiny is voiced by a growing number of researchers who highlight the ineffectiveness of the move to ethical AI principles (e.g., Lauer 2021). For instance, Mittelstadt (2019) emphasises that ethical AI frameworks result in "vague, high-level principles, and value statements which promise to be action-guiding, but in practice provide few specific recommendations and fail to address fundamental normative and political tensions embedded in key concepts" (p. 501). This statement reflects upon two main criticisms regarding the proliferation and use of ethical AI frameworks put forward.

First, ethical AI principles lack consensus over their operationalisation. The issue lies in translating normative concepts, such as autonomy and justice, into technical rules and best practices adopted by AI practitioners. Mittelstadt (2019) argues that these concepts are significantly shaped by local contexts, and high level abstractions may not provide adequate guidance. In this regard, practitioners often translate principles and make normative decisions in the way they deem fit in the absence of coherent implementation, which entails that the issue of operationalising principles "is kicked down the road like the proverbial can" (Mittelstadt 2019, p. 503).

Second, tensions emerge when ethical AI principles are adopted in specific contexts. These tensions may reflect the presence of a moral trade-off, wherein two goals or values conflict with each other and one cannot be pursued without foregoing the other. For instance, a report by the AI Committee of the UK House of Lords (2018) mentions that "it is not acceptable to deploy any artificial intelligence system which could have a substantial impact on an individual's life, unless it can generate a full and satisfactory explanation for the decisions it will take" (n.p.). Here a tension lies between the use of algorithms for social benefit and assuring understandings of the algorithms to the wider public. As such, without acknowledging the impact of ethical tensions, Whittlestone et al. (2019) argue that standards adopted may be unachievable and regulations which ought to protect certain values may inadvertently impact others.

1.2 Motivation for Research

Thenceforth, although AI is increasingly used to enhance the efficiency and reduce costs of PSOs, decisions made by such organizations have great impacts on both individuals and society. Despite this, Hickok (2021) emphasises that a limited number of stakeholders still remain in charge of making crucial decisions regarding AI, such as deciding on what is prioritised, where AI systems will be used and which decisions they deem important. This marks an asymmetry between those who decide and the wider public impacted by these decisions, leading to a power asymmetry between involved stakeholders.

Thus, this research follows the argumentation of various ethical guidelines that posit citizens must be more involved in crucial decisions regarding AI (e.g., Floridi et al. 2018), as doing so would enable them to hold decision-makers accountable regarding the societal impacts of AI. The plurality of opinions and diverse viewpoints may then be captured in these systems, which are designed with human-centricity and beneficence in mind.

The Netherlands is taken as a case study in this research. This serves as an interesting case since despite the country showing large progress in adopting AI systems in the public sector and becoming a recent frontier on the ethical use of AI (Asser Institute 2021), over 60% of Dutch citizens distrust the national government (Statista 2022). This is paradoxical considering that trust is regarded as an important factor underpinning the use and social acceptance of AI in the public sector, especially when values are leading (Chen et al. 2021). As other European countries start to engage with and implement AI projects throughout their public sector, insights from the Netherlands may provide ample knowledge on how to roll out such projects whilst ensuring that citizens' needs are considered. Hence, the main research question and sub-questions addressed in this research are:

Main RQ:	How do Dutch citizens perceive the ethical implications of AI use in the public sector?
Sub-Q One:	What are Dutch citizens' understandings of ethical AI principles?
Sub-Q Two:	What values are prioritised by Dutch citizens when tensions emerge following the implementation of ethical AI principles?

The aforementioned questions are answered by conducting focus groups with Dutch citizens to grasp how they make sense of ethical AI principles in practice – relating to the operationalisation element – and what values are leading within specific local contexts. The latter provides insight on addressing tensions resulting from the implementation of ethical AI principles. Although the Dutch citizens questioned are not meant to be representative of the broader society, they can provide initial exploratory insights into how citizens perceive these elements, which has thus far been under researched in academic literature (Ingrams et al. 2022).

The transcripts of the focus groups are deductively coded using the aggregate of the frameworks from Floridi et al. (2018) and Ashok et al. (2022) – the Extended AI4People Framework. This serves as the theoretical foundation to which citizens' understandings of the ethical implications of AI use are compared to, as the Framework discusses established guidelines, practices and research in the AI domain. Additionally, the four main ethical tensions identified by Whittlestone et al. (2019) are analysed to grasp what values citizens prioritise when the ethical tensions emerge. Hence, combining both the Extended AI4People Framework and view on the ethical tensions provides a guiding lens to highlight data that aids in answering the research questions posed, generating greater insight about how citizens perceive the technological changes in the public sector and their associated repercussions.

By moving beyond citizens' ethical perceptions of AI as a mere obstacle that needs to be overcome, this research contributes to understandings of how Dutch citizens perceive the use of AI in their society, alongside the ethical implications such use entails. A single case-study allows greater insight on the intertwinement of technology, local norms and the context of use, which are deeply embedded within the cultural and social context. Additionally, highlighting the values that Dutch citizens deem important can aid practitioners when resolving trade-offs which emerge when ethical AI principles are implemented in practice. As there is not one best method to minimise tensions, Abedin (2021) highlights that balancing organizational, environmental and individual needs is key when deploying ethical principles to ensure that trust and approval in AI systems are maintained. By addressing these points, Berendt (2019) argues that such insights can aid

developers with better designing ethical AI systems suited towards the end users' needs – the citizen.

1.3 Structure of the Paper

The structure of this research is as follows; the next section discusses key research in the domain of AI, including advancing a working definition and application areas. The literature background also emphasises the roles and perspectives of citizens' on AI, established ethical frameworks and principles, alongside relevant critiques. In particular, the main theoretical lens – the Extended AI4People Framework – is presented alongside the four main ethical tensions, which serve as the foundation of this research. The research design then follows, wherein the focus group method adopted is elaborated upon, alongside relevant practical elements of how this method is conducted and utilised.

Further, the results convey Dutch citizens' perceptions on the ethical implications of AI use in the public sector, and are reported under three main themes. These themes coincide with the research questions posed in this research, hence providing a concrete structure to the results. The next section advances discussions about the results of this research, highlighting novel perspectives put forward by citizens and relevant contextual elements. In the conclusion, the research questions are explicitly answered, both academic and practical implications are noted, alongside the limitations of this research and suggestions for future research.

2 Literature Background

This section provides an overview of AI, with a particular emphasis on *ethics*. First, AI is defined due to its conceptual ambiguity and its application areas in the public sector are highlighted. Following this, the benefits and values of AI as perceived by citizens are discussed. This shifts the focus to the *ethics* aspect of AI, wherein ethical AI frameworks and principles are elaborated upon. The Extended AI4People Framework is presented, since it serves as the main theoretical background for this research. To provide a more nuanced view, critiques against the framework are also advanced and considered, highlighting current shortcomings to ethical approaches. Finally, the citizens' importance in AI development is reiterated, and provides a segue to the results wherein citizens' perceptions on the ethical implications of AI are discussed.

2.1 Overview of AI in the Public Sector

The discipline and field of AI has been discussed for decades by both practitioners and academics alike, but only recently gained relevance for and momentum in the public sector. This marks a critical step considering that the public sector is perceived as playing a crucial role in the development of AI – both in advancing legislation (e.g., Misuraca and van Noordt 2020) and incorporating it within application areas (e.g., Wirtz et al. 2019). Nevertheless, research on the applications of AI in the public sector remain limited despite the increasing momentum, which in turn brought about confusion regarding how AI is defined and understood (Ahn and Chen 2022).

Thus, Maragno et al. (2022) argue that the present problem is twofold. First, AI is still being used as a general term encompassing diverse technologies such as video recognition and machine learning. Second, ambiguity remains regarding the role governments are playing within applications of AI, the decisions being made and reporting on successful implementation. With the latter, Misuraca et al. (2020) are seen as having conducted one of the few research that focuses on AI projects within the European setting.

2.1.1 Defining AI

Despite that the field of AI has been studied for several decades, a consensus regarding how to define the term is yet to be achieved. Legg and Hutter (2006) argue that the ambiguity limits our understandings of AI, which they perceive as a problem that persists nowadays. A recent report published by AI Watch reiterates this issue, thus attempting to put forward an operational definition of AI (Samoili et al. 2021). This highlights how attempts regarding defining AI still persist, adding to the conceptual ambiguity.

Previous definitions of AI refer to the systems in relation to people, and were thus disproportionately human-focused. For instance, Rich et al. (2009) defined AI as "the study of how to make computer do things, which at the moment, people do better" (p.3). Similarly, Russel and Norvig (2010) put forward the definition that AI can be organised into four categories: "systems that think like humans; systems that act like humans; systems that think rationally; systems that act rationally" (p. 2). These definitions are then followed by ones describing AI systems more distinctly in relation to people. Adams et al. (2012) is an example of such definition, who describes AI as a "systems that could learn, replicate, and possibly exceed human-level performance in the full breadth of cognitive and intellectual abilities" (p. 28).

More recent approaches to defining AI focus on defining these systems as *intelligent* beings, beyond reference to humans. For example, Wirtz et al. (2019) suggest to first delineate and define the term *intelligence* explicitly, which can then be applied to machines and provide a more nuanced definition to the compounded *artificial intelligence*. Accordingly, Legg and Hutter (2006) define intelligence as possessing the capabilities to interact with and acquire information pertaining to past experiences, alongside handling uncertainty. Artificial is then defined by Patrick and Fattu (1986) as a copy of something natural, which is produced by humans.

The aforementioned definitions thus highlight that the core characteristics of AI include a machine-based system which displays intelligent human-like behaviour. Such behaviour includes perception, learning, and understanding. These allow AI to mimic human thinking and practices when targeting efficient solutions, leading to better performances (Wirtz et al. 2019). The working definition put forward by the EC (2019a) HLEG AI highlights the elements posed above, and is thus adopted in this research – AI is seen as a "systems that display intelligent behaviour by analysing their environment and taking actions – with some degree of autonomy – to achieve specific goals" (p. 1).

2.1.2 Application Areas

As a result of increasing societal pressures to digitalise their processes, Pfotenhauer and Jasanoff (2017) indicate that governments responded by catering to innovation discourses as to appear politically legitimate and economically feasible. Consequently, governmental offices and PSOs worldwide are implementing AI applications. Research initially focused on AI's ability to deliver enhanced outcomes to the private sector, including intelligent automation, virtualisation of labour and complimentary benefits to the abilities of personnel (Bataller and Harris 2016). Hence, the public sector is able to use insights and best practices from the private sector wherein AI uses are perceived as

common knowledge, whilst safeguarding public interests and upholding the provision of public values (Wirtz et al. 2019).

De Sousa et al. (2019) conducted a systematic literature review (SLR) to identify studies published during the 21st century with a focus on AI in the public sector. The 59 resulting studies discussed AI applications in nine of the ten functions of government, as ascertained by the Classification of Functions of Government. Accordingly, the functions of government where AI use is most reported are: general public services, economic affairs, environmental protection, public order and safety, housing and community amenities, social protection, health, defence and finally, education. The only function of government where the authors did not find research regarding AI applications is within the recreation, culture and religion domain.

Diving deeper into some of the AI application areas, various authors refer to education as a function of government where AI has a pivotal role. Within this context, following their SLR on AI use cases in education between 2010-2020, Chen et al. (2020) argue that AI has various beneficial use cases. These include aiding in the assessment of students and the grading of exams or reports, providing personalised intelligent teaching, and further enabling online education. The AI techniques associated with these benefits include adaptive learning methods, image recognition and prediction systems, learning analytics and virtual personalised assistants respectively.

Nevertheless, some challenges from using AI within this domain persist. These include the changing role of teacher vis-à-vis their students in the classroom. From the teacher's perspective, they may either wholly resist AI implementations due to their beliefs regarding own self-competency, or due to the overreliance on AI and associated unrealistic expectations that are often unmet (Kessler 2018). From the student's perspective, AI seems to provide the relevant tools required to produce outcomes without the need for knowledge processing work, undermining students' learning (Zhai et al. 2021). For instance, instead of exploring relevant examples to questions posed by teachers, students can obtain results using smart tools, compromising their learning.

Also considered a vital area where AI plays a positive role is the domain of public order and safety. For instance, various countries such as the Netherlands implemented intelligent image processing software to enhance crowd monitoring and control. This software, which is also considered a facial recognition system, is supported and used by various intelligence agencies and police departments to identify and locate criminals within the natural environment, missing persons and victims (Misuraca et al. 2020).

Besides the Netherlands, other countries such as China and the UK employ edge video analytics as an essential means of upholding public safety. This method involves the gathering and analysis of data continuously from live camera streams or sensors. Zhang et al. (2019) highlight how video analytics utilise AI to address the so-called four W Problem: the actor (Who), doing something (What), the place it occurs in (Where), and the time (When). The answers to these questions are relevant within the contexts of building awareness about situations, providing alerts when something goes wrong and for detecting relevant objects (e.g., missing vehicles and suspects).

Local governments also incorporate AI within their structures in order to enhance interactions between their organizations and the wider public (e.g., businesses and citizens). Tangi et al. (2020) argue that such technology reinvents communication lines between PSOs and their constituents, making it easier for public demands to be expressed and heard. For instance, a report published by the European Regional Development Fund shows how various Dutch cities are using chatbots to answer questions posed by the public as the next step in digital customer service, thereby reducing administrative burden on the staff (de Voogd 2019).

Van Noordt and Misuraca (2019) highlight that the current application of chatbots is limited to the emergence of only first-order changes – those pertaining to the automation of established activities and slight changes to the organization as to facilitate the introduction of chatbots. This denotes the lack of second-order changes within chatbot applications in the public sector – those pertaining to the radical transformation of public service delivery, changes in governance mechanisms, or new methods for citizen engagement surpassing established figures. Despite this, the improved communication between governments and citizens seen in first-order changes still provide value by easing requirements of looking for information (Aoki 2020). The resultant response uniformity, quality and timeliness leads to strengthened public trust in AI systems, and thus such changes should not be neglected.

As such, the aforementioned AI application areas and use cases are not meant to provide an exhaustive account. Instead, they show the increasing implementation of AI in the public sector. Some challenges still remain and the full potential of AI in various cases is yet to be met, but the overview shows how the implementation of new technologies in the public sector is developing, which for long was considered lagging behind the private sector in this regard.

2.2 Instrumental and Value-Based Dimensions

The application areas discussed above show how AI permeates expansive areas in citizens' daily lives, including education, public safety and their interactions with local government. Despite this, König (2022) highlights that evidence on citizens' perceptions of AI remain limited and disparate. Whereas Marcinkowski et al. (2020) put forward findings showing that citizens are sceptical of AI use in political decision-making, Miller and Keiser (2021) show citizen acceptance of AI use in simple public administration tasks. This highlights the existence of disparate opinions on how AI in the public sector is viewed by citizens, requiring further analyses.

Hence, to better understand how citizens perceive and make sense of AI use in the public sector, Ingrams et al. (2022) argue that discussions surrounding improvements offered by AI focus on both instrumental impacts of AI use, alongside the moral and societal outcomes of AI. The former iterates the efficiency advances associated with AI and is referred to as the instrumental dimension, wherein the latter highlights the impact of AI on both the generation of public values and public sector-citizen relationships, and is dubbed the value-based dimension.

The instrumental dimension focuses on the perceived positive benefits of AI use in the public sector due to technical advances that improve the efficiency of PSOs and their respective services. This line of argumentation is the most cited in research when referencing rationales behind the implementations of AI projects by the public sector – in pursuit of efficiency while lowering costs. These benefits can be achieved through various ways.

First, AI offers benefits through decreasing both indirect and direct costs. For instance, Chun (2008) shows how AI is used in optimising immigration forms through improving workflow case assignments, providing decision support and follow-up actions, alongside learning from established practices. Such actions have minimised employees' workloads and improved the overall flow of work, entailing both efficiency and economic benefits. Meijer and Wessels (2019) also highlight how AI use in predictive policing allows law enforcement to better identify problematic geographical areas that aids in better deploying existing resources. In this regard, public employees can focus their efforts on specialised activities and tasks, allowing machines to support routine procedures.

Second, AI may be used to enhance decision-making accuracy. In a research conducted by Nasseef et al. (2022), the ability of AI to enhance the quality of decisions within public healthcare is empirically shown by using Saudi Arabia as a case study. Similarly, research by Moingeon et al. (2022) emphasises how AI can be converged with existing health

technology to provide personalised therapy and preventative measures. Combining information about patients' physiology, exposure to environmental risks and disease features, AI aids in decision-making through the provision of predictive models. Hence, though the results are limited to the health sector, the perceived benefits of AI in decision-making are gaining momentum in research, showing AI's potential in diverse application areas (Kuziemski and Misuraca 2020).

Although the instrumental impacts of AI use are largely positively perceived in academic research and by public administrations, the emphasis of the value-based dimension on public values entails that more contentious elements are at play. This dimension focuses on citizens' views related to the social, normative, and political implications of AI use. In the public sector, AI use is often promoted under the guise of objectivity and neutrality. Nevertheless, this is not the case in practice as Kitchin (2017) argues that AI is "created for purposes that are often far from neutral: to create value and capital; to nudge behaviour and structure preferences in a certain way; and to identify, sort and classify people" (p. 18). Worries thus emerge about AI's ability to replicate deep, structural biases found in society that are embedded into computer codes, leading to ill-outcomes and biased decisions (e.g., Miller and Keiser 2021).

Accordingly, Kieslich et al. (2022) advance the four most prominent concerns voiced by citizens regarding AI use. These concerns contradict the core public sector ethos, which is based on the Weberian principles regarding transparency, oversight, equality and upholding the public's well-being (Willems et al. 2022). Accordingly, identifying these principles helps shed greater light on what citizens find problematic with AI implementation and its ethical implications, providing a starting point on what elements can be improved to enhance citizens' perceptions.

First, citizens fear that AI use may threaten and violate their privacy. Citizens hold concerns that their personal data is collected and processed without their consent or in accordance with established laws (Wirtz et al. 2019). Rössler (2004) argues that such violations can occur in one of three ways: "as illicit interference in one's actions, as illicit surveillance, [or] as illicit intrusions in rooms or dwellings" (p. 9), which Calo (2011) argues are applicable within an AI context.

Willems et al. (2022) discuss privacy concerns associated with AI more concretely. The authors argue that the self-learning algorithm, the foundation for AI, requires large data sets that are based on personal data as inputs to the systems, impeding on citizens' privacy. The use of these algorithms and the input data collected are often deployed in a non-transparent manner without a definitive ownership structure, adding to the complexity of citizens establishing when their information is accessed or the purpose

behind using their data. This in turn can hamper public sector AI projects since citizen support is considered a crucial precondition for justifying the use of new technologies (Chen et al. 2021).

Second, citizens worry about the emergence of unfair outcomes when AI is used, particularly in decision-making areas. Issues regarding fairness are discussed in literature, woth AI outcomes adversely impacting certain societal groups including the elderly (Roseman and Stephenson 2005) and low-income members (Zhou et al. 2022). Starke and Lünich (2020) argue that fairness is a vital criterion for evaluating AI systems, wherein systems that are perceived as a potential cause of detrimental consequences on the implementing institutions.

One method suggested in research to address issues regarding fairness is put forward by Nakao et al. (2022). They suggest incorporating citizens as providers of feedback to AI systems. More concretely, AI system designs can encompass understandable and interactive *human-in-the-loop* interfaces, which enable ordinary citizens without technical expertise to spot issues pertaining to fairness and be able to fix them. In an implemented prototype, the authors show how the feedback mechanism works. Citizens can view why certain predictions are put forward, and can then change the weight of each feature assessed as a way to enhance fairness. Conceptions of fairness are not static worldwide, and thus cultural dimensions are explored in each setting to see what citizens regard as fair, which is then incorporated within the human-in-the-loop approach.

Third, removing the human element from decision-making is negatively perceived by citizens who believe that certain areas are more suited for human evaluations. This is due to the presence of immeasurable characteristics (e.g., the role of human empathy on decisions). Starke and Lünich (2020) argue that decision-making processes which incorporate some degree of human oversight are more positively perceived by citizens as they are regarded as fairer and more legitimate.

Nevertheless, concerns regarding black box decision-making influenced the emergence of contemporary data protection regulations and laws regarding the incorporation of citizens' rights to intervention when decision-making occurs through AI support (Almada 2019). These interventions, which can either be perceived as a minimum requirement for data processing or a guiding norm for AI-aided decisions, are a mechanism to ensure safeguarding citizens' freedoms, rights, and interests. Nevertheless, two main issues exist within this approach that ought to be considered. These issues include questions regarding which decisions should be able to warrant interventions (e.g., *only fully automated decisions or is partial automation enough?*), and the lack of knowledge citizens have to

exercise their rights fully. The presence of these issues indicates that concerns regarding black box decision-making still persist nowadays.

Finally, citizens fear that AI cannot replicate elements pertaining to human complexity, wherein AI systems fail to address unique aspects in individual cases and can lead to inaccurate decisions. Empirical evidence put forward by Dietvorst et al. (2015) shows that citizens negatively perceive and refrain from engaging with algorithms that have made mistakes, and even lose trust in these systems.

A different perspective is found in research put forward by Jarrahi (2018), who posits that both AI and humans complement each other and have their own strengths that impact decision-making processes. These processes are often defined by their complexity and uncertainty. From the side of the AI, the systems provide an analytical approach to decision-making, compounded with sophisticated computational information processing capacities. On the human side, people can offer more comprehensive, intuitive approaches in the face of complexity and uncertainty, addressing concerns put forward by citizens. Hence, rather than replacing human contributions, AI systems should be regarded as augmenting human capacities than replacing it. Jarrahi (2018) refers to this as the *Human-AI symbiosis* in organizational decision-making.

As such, the abovementioned instrumental and value-based dimensions provide a guiding lens to understand how citizens perceive the use of AI in the public sector. Within the domain of AI, the citizen view is construed as important since AI systems are considered socio-technical artefacts. This entails that AI is perceived to incorporate more than just technical elements, moving beyond its "encoded procedures for transforming input data into a desired output, based on specified calculations" (Gillespie et al. 2014, p. 1). Instead, AI systems are viewed as being embedded in a certain environment impacted by institutional and societal structures. Kitchin (2017) argues that AI "can be thought about in a number of ways: technically, computationally, mathematically, politically, culturally, economically, contextually, materially, philosophically, ethically and so on" (p. 16). Consequently, Lee (2018) highlights that understanding citizens' views on this system is important, since how they understand AI and its functions aid in their acceptance.

Thus, the use of AI is not only driven by output optimisation, but also incorporates elements that safeguard important societal values. In this regard, some researchers argue that on the input side, AI should consider citizens' sociocultural complexities while providing understandable explanations on the output side for the general public (Riedl 2019). In the same vein, Gurr (1971) highlights that "governance can be considered legitimate in so far as its subjects regard it as proper and deserving of support" (p. 185).

For citizens, this entails that AI-related risks are adequately handled, alongside safeguards that uphold their safety and stability within their environment.

2.3 Ethical AI Framework(s)

Benefits of AI use in the public sector are increasingly covered in research (e.g., de Sousa et al. 2019), but questions remain regarding their ethical and responsible use, alongside their immediate and far-term consequences on societies, including citizens (Leikas et al. 2019). When in use, AI systems are progressively engaged in situations where their outcomes are judged as either morally good or bad, based on their imposed effects on society. As such, ensuring the emergence of socially preferable outcomes requires a balance between the benefits AI offers and minimising potential negative consequences. An ethical approach to AI thus upholds the notion that compliance with existing rules regulations, though necessary, is an insufficient standard as it is not the most that can be done (Floridi 2018).

Floridi et al. (2018) highlight that adhering to an ethical approach provides a *dual advantage*. On the one hand, adopting organizations can reap the benefits of leveraging new opportunities enabled by AI that are preferred by the public, due to the potential of enhanced services and cost savings. This provides social value from which the organizations can benefit. On the other hand, the adopting organizations also minimise risk and the emergence of costly mistakes by following an ethical approach. Without this approach, the negative outcomes of AI are deemed socially unacceptable as not enough was done to mitigate associated risks, and the systems are rejected. In this regard, Floridi et al. (2018) highlight that the adoption of AI in the public sector can only occur under conditions where the benefits are seen as important, and risks are minimised through the adoption of risk management approaches.

2.3.1 AI4People Framework

The strong interrelation between social acceptability and incorporation of ethics in technology paved the way for the development of ethical AI frameworks (Leikas et al. 2019). Floridi (2018) argued that the field lacked a direction and collective vision that would enable AI to surpass its rate of development. Although several ethical AI principles were adopted at the national level (e.g., in the UK) or published by organizations (e.g., The Tenets of the Partnership on AI), no overarching ethical AI framework was published. In this regard, Floridi et al. (2018) introduced an ethical AI framework – the AI4People Framework – which was established after a comprehensive synthesis of existing principles published by six trusted sources. These sources are:

- The **Asilomar AI Principles** (2017), sponsored by the Future of Life Institute, in association with participants of the Asilomar conference
- The **Montreal Declaration for Responsible AI** (2017), developed by the University of Montreal, ensuing the conference on the Socially Responsible Development of AI
- The **General Principles** put forward in the Ethically Aligned Design: A Vision for Prioritising Human Well-being with Autonomous and Intelligent Systems (2017) by the Institute of Electrical and Electronics Engineers
- The **Ethical Principles** discussed in the Statement on Artificial Intelligence, Robotics and 'Autonomous' Systems (2018), published by the EC's Group on Ethics in Science and New Technologies
- The **Five Overarching Principles for an AI Code** stated in the report AI in the UK: ready, willing and able? (2018) by the UK House of Lords Artificial Intelligence Committee
- The **Tenets of the Partnership on AI** (2018), a cross-sector organization encompassing academics, researchers, civil society organizations, and others

The authors indicate that the framework represents a European approach to ethical AI (Floridi et al. 2018). Five main ethical AI principles are thus incorporated within the framework, which were the most salient during the synthesis of the sources. The resulting principles are: beneficence, non-maleficence, autonomy, justice and explicability.

The first four principles overlap with the core principles of bioethics. Floridi (2013) suggests that this result is not surprising since bioethics is closely intertwined with digital ethics, as both incorporate "new forms of agents, patients, and environments" (Floridi et al. 2018, p. 696). Nevertheless, the final principle of explicability, which is not part of the bioethics principles, is deemed necessary within the context of AI and was thus added. Explicability is regarded as incorporating both intelligibility and accountability.

2.3.1.1 Beneficence

Beneficence necessitates that AI should be developed and deployed to empower the common good. This includes promoting the well-being of humans and upholding their basic rights. Despite the fact that beneficence is the most prominent principle among the six sources, the principle is referred to through varying degrees throughout. For instance, while the Asilomar AI Principles (2017) promote the view that AI should strive towards

achieving benefits for humanity, the Montreal Principles (2017) refer to AI benefiting not only humans, but all sentient creatures. On the other hand, the EC's (2018) definition encompasses both human dignity and sustainability. Floridi et al. (2018) regard this as the broadest understanding of beneficence between the analysed sources.

Sustainability, as an underlying component in the benefice principle, entails that AI "must be in line with the human responsibility to ensure the basic preconditions for life on our planet, continued prospering for mankind and preservation of a good environment for future generations" (EC 2018, p. 19). The incorporation of sustainability can be perceived as promoting an information systems' perspective, which requires organizations to consider aspects such as the environment (Thiebes et al. 2021).

McKnight et al. (2002) highlight that beneficence is aligned with beliefs regarding helpfulness, benevolence and acting with a purpose. Accordingly, AI systems working under this principle ought to act in the citizens' interests. This includes maximising benefits when possible, and minimising manipulative or opportunistic actions relative to humans, other sentient beings and the environment.

2.3.1.2 Non-maleficence

Whereas the beneficence principle promotes AI use for the common good, non-maleficence entails that AI operates with the intent of *do no harm*. The over or misuse of AI may result in negative consequences, most notably violations of personal privacy. Safeguarding individuals' privacy is interlinked with their access and control over their personal data, alongside how it is used. Other mentioned negative consequences include an AI arms race threat or the ever-evolving self-development of AI. In this regard, this principle cautions against and calls for upper limits on the capabilities of AI. Hence, non-maleficence is based on assumptions of reliability and integrity, where AI systems work in an honest manner and produce consistent outcomes while adhering to ethical principles embedded in their design (Thiebes et al. 2021).

Research proposes various methods for protecting citizens' privacy both within the training and operating of AI. This includes adding noise to data, thereby expanding the size of training datasets and minimising the memorising of training samples (Song et al. 2013). Alongside this, the use of relevant and credible execution environments is suggested, with the environments representing a particular execution platform such as a database management system or operating system (Zhu et al. 2020). These approaches are encouraged due to the sensitivity of citizens' stored information and their concerns regarding privacy.

Nevertheless, Floridi et al. (2018) discuss some confusion regarding this principle. It remains unclear whether *do no harm* applies on the side of the AI developers where intent plays a role or the AI systems themselves and their potentially unpredictable behaviour. Moor (2006) offers a possible answer to this conundrum when suggesting three possible methods to educate AI systems to be ethical: training AI into *implicit ethical agents* by restraining their actions to prevent unethical outcomes, *explicit ethical agents* wherein some actions are explicitly approved while others are prohibited, and *full ethical agents* wherein AI has consciousness and free will.

Wang and Siau (2018) argue that the explicit ethical agent is most salient in academic research and deemed practical. In this regard, more emphasis is placed on the AI developers and the conditions they set for AI systems. This may be the case until "the ultimate goal of machine ethics" is achieved, which encompasses "creating a machine that itself follows an ideal ethical principle or set of principles" (Anderson and Anderson 2007, p.15).

2.3.1.3 Autonomy

In bioethics, the third principle of autonomy incorporates the assumption that individuals retain the right to decide on matters pertaining to any treatment they receive. That is referred to as informed consent, wherein patients are given information about their course of action and contemplate between various options (Farrell et al. 2014). This principle is violated when patients do not possess the required mental faculties to uphold decisions promoting their best interests, hence autonomy is forcedly ceded (Floridi et al. 2018).

Within an AI context, this principle becomes complex in the absence of a definite definition. Autonomy is elaborated upon in two main ways by the six sources – as promoting human agency and autonomy, or as restricting the autonomy of AI systems. The former highlights that the use of AI systems presupposes that individuals cede some autonomy and power. Hence, autonomy as a principle within an AI context aims to balance power relations between agents and that delegated to machines concerned with decision-making. This sentiment is echoed by the Asilomar AI Principles (2017) when stating that "humans should choose how and whether to delegate decisions to AI systems, to accomplish human-chosen objectives" (n.p.) and by the EC (2018) wherein AI "must not impair the freedom of human beings to set their own standards and norms and be able to live according to them" (p. 16).

The latter emphasises that the principle of autonomy does not only set out to safeguard the autonomy of individuals, but also ensures restrictions can be placed on that of the machines'. Any autonomy ceded to AI must be reversible. Floridi et al. (2018) dub this

as the *meta-autonomy* model of AI. This model states that agents must always be able to control and have the freedom to dictate which decisions are within their prerogative, and when decision-making powers can be ceded in case efficiency gains and benefits are anticipated. As these decisions do not exist in a vacuum, social and cultural contexts impact the division of autonomy between both parties. Hence, safeguarding autonomy necessitates that integrity and reliability risks are mitigated through securing a balance within machine-led and human-led decision-making processes (Thiebes et al. 2021).

2.3.1.4 Justice

As the last core bioethics principles, justice within a healthcare setting is related to resource distribution across the groups in society. The AI context is similar in this, as the principle encompasses the notion that AI should enable phasing out all types of discrimination, resulting in both shared prosperity and benefits in society (Asilomar AI Principles 2017). This highlights that justice is not understood in the judicial sense of following regulations and laws, but from an ethical perspective wherein elements regarding fairness are emphasised (Floridi et al. 2018). The Montreal Declaration (2017) reiterates the ethical, rather than judicial, perspective when stating that the "development and use of [AI] must contribute to the creation of a just and equitable society" (p. 13). Alongside the outcomes of AI being non-discriminatory, the principle also calls for caution against introducing biased datasets for training.

Comparable to the non-maleficence principle, justice within AI aligns with beliefs regarding helpfulness, benevolence and acting with a purpose (Thiebes et al. 2021). Contemporary research regarding the justice principle in AI highlights the presence of biases in AI systems alongside ways to overcome them (e.g., Mehrabi et al. 2021), as well as developing approaches to quantify fairness in such systems (e.g., Bellamy et al. 2019). In short, Floridi et al. (2018) summarise the main issues that the principle encompasses as the following: (1) AI is used to eliminate past mistakes, such as bias; (2) AI is used to create societal benefits, that are shared by all; and (3) AI does not contribute to new harms, such as eroding established structures.

2.3.1.5 Explicability

Explicability is the only principle falling outside the realm of bioethics and is deemed necessary to incorporate by Floridi et al. (2018) within the ethical AI principles. The need to incorporate this principle is the result of issues regarding the development of AI. Only a small group of technical experts are currently part of the design and development of AI systems, with these systems having profound impact on peoples' daily lives.

Explicability is referred to by various synonyms in the six sources, including accountability, understandability, transparency and interpretability. Despite the different terminology, Floridi et al. (2018) argue that the underlying meaning is the same – how AI functions is often not visible or understandable to the wider public who are impacted by the outcomes of AI systems. In this regard, explicability incorporates both an epistemological element and an ethical element. The former encompasses the concept of intelligibility, which relates to questions regarding *how does AI work?* In this regard, intelligibility focuses on the greater production of understandable AI models that perform well and are accurate. The latter embraces the concept of accountability, which relate to questions regarding *who is responsible for how AI works?* Hence, the principle highlights elements pertaining to explainable and intelligible performance, functionality and competence of AI systems (Thiebes et al. 2021).

In contemporary AI research, explicability is a common theme discussed due to the perceived black box decision-making surrounding AI systems – that they are not transparent nor accountable. These elements have a negative effect on citizens' perceptions since they do not understand nor trust the outcomes of AI systems (Thiebes et al. 2021). Thus, various approaches are discussed and analysed in literature to overcome issues pertaining to explicability, such as the production of transparent and understandable models, alongside instituting post-hoc explainability (Arrieta et al., 2020). Other researchers promote attempts to quantify uncertainties (e.g., Bellamy et al. 2019) and auditing of AI systems (e.g., Mökander and Floridi 2021). From an information systems' perspective, promoting explicability goes beyond ensuring AI compliance requirements, as it is considered key for AI acceptance by citizens and public administrators.

2.3.1.6 Interlinkages between Principles

Floridi et al. (2018) argue that the ethical AI principles enable people to understand the relationship between themselves and the disruptive technology in an intelligible manner. Explicability is considered an enabling principle for the first four ethical AI principles, which are the core bioethics principles. The interlinkage between explicability and the first four principles are highlighted by Floridi et al. (2018) in the following:

> For AI to be beneficent and non-maleficent, we must be able to understand the good or harm it is actually doing to society, and in which ways; for AI to promote and not constrain human autonomy, our "decision about who should decide" must be informed by knowledge of how AI would act instead of us; and for AI to be just, we must ensure that the technology— or, more accurately, the people and organizations developing and deploying it—are

> held accountable in the event of a negative outcome, which would require in turn some understanding of why this outcome arose (p. 700).

As such, the five ethical AI principles are proposed as constituting a comprehensive ethical AI framework, which helps adopting organizations make use of the *dual advantage* of an ethics approach. Floridi et al. (2018) underscore that these principles must be incorporated within established practices regarding AI. More specifically, the design and development of AI should align with the goals of promoting well-being and reducing inequalities, while upholding autonomy and ensuring that benefits gained are equally distributed within the society. The explicability of AI is also crucial, as it is the mechanism by which citizens' trust can be built (Thiebes et al. 2021). Floridi et al. (2018) promote this view and recommend public administrators to establish guidance for AI developers, users and rule-makers, as such a multistakeholder approach is considered a valuable method to serving societal needs.

2.3.2 An Extended AI4People Framework

Despite the salience of the five ethical AI principles advanced by Floridi et al. (2018) between the analysed sources, Ashok et al. (2022) argue that the principles do not share an ontological basis, which may hinder ethics evaluations and lead to discrepancies. These discrepancies occur amongst actors, fuelled by economic and social motives, and their interpretations of the AI principles in practice. Accordingly, the authors provide an extension to the AI4People Framework proposed by Floridi et al. (2018), which is dubbed in this research as the Extended AI4People Framework. The extension of the framework is done in two main steps.

First, Ashok et al. (2022) incorporate an ontological framework to further categorise the ethical AI principles, providing an ontological foundation for the principles and reducing ambiguity due to their abstract nature. Following this, the authors develop a list of ethical implications pertaining to each of the ethical AI principles, as discussed in AI ethics literature. The resultant framework – the Extended AI4People Framework – thus categorises the ethical AI principles on an ontological framework and integrates the ethical implications for the principles.

2.3.2.1 Ontological Framework

First, rather than defining ontology from a philosophical viewpoint, an information systems' perspective is put forward instead, highlighting that ontology aims to produce a "shared taxonomy of entities" (Smith and Welty 2001, p. vi). Following this definition, the authors utilise three main ontological frameworks (Popper 1979; Ogden and Richards

1989; Denning and Rosenbloom 2009), which aid in the identification of three domain areas to discern the relationship between the use of AI and its ethical impacts.

The first domain is the physical domain, referred to as Popper's World I ontology, the referent/object by Ogden and Richards (1989), and the physical by Denning and Rosenbloom (2009). The second domain is the cognitive domain, referred to as Popper's World II ontology, the reference/interpretant, and the social respectively. The third domain is the information domain, referred to as Popper's World III ontology, the symbol/sign, and the life respectively. Table 1 shows an overview of the three domains, alongside how they are referenced by the respective ontological framework authors. Ashok et al. (2022), building on research by Liu (2000), introduced a fourth domain, which they title governance. This domain aims to capture elements related to the functioning of the information system.

Table 1 - Ontological Framework

Ontological Framework	**Popper (1979)**	**Ogden and Richards (1923)**	**Denning and Rosenbloom (2009)**
Physical domain	World I	Referent/object	Physical
Cognitive domain	World II	Reference/interpretant	Social
Information domain	World III	Symbol/sign	Life

2.3.2.2 Ethical Implications

Second, Ashok et al. (2022) conducted a SLR, alongside qualitative synthesis to discern the ethical implications of AI use discussed in academic research, until a point of theoretical saturation is reached. The inclusion criteria is peer-reviewed studies discussing AI applications and ethical implications in English, published after 2000 as to encompass current AI research. From 195 initial identified records, 59 papers remained as they fit the inclusion criteria (for an overview, see Ashok et al. 2022, p. 15).

Following the SLR, a qualitative synthesis was done on the remaining papers through the use of template analysis. Such an analysis allows the emergence of first conceptual themes, which can then be grouped into constituent themes and later identify global themes. An a-priori template was used, which outlined the ethical AI principles advanced by Floridi et al. (2018). Based on the template, the ethical implications of AI use discussed in the literature review papers were coded, and further categorised into global themes. The coding took place over multiple rounds as a method for the authors to check for reflexivity.

2.3.2.3 The Extended Framework

The combination of the ontological framework and the use of the a-priori template based on the ethical AI principle yields an extension to the AI4People Framework. The Extended AI4People Framework shows both the principles and their ethical implications clustered around the four main ontological domains – physical, cognitive, information and governance – representing the global themes. Alongside discussions of the ethical implications of each of the five principles, Ashok et al. (2022) put forward a sixth principle pertaining to governance, which they argue is relevant for the digital technology environment nowadays. Table 2 shows an overview of the Extended AI4People Framework.

Table 2 - Extended AI4People Framework

Ontological Domain	AI Principle	Ethical Implication	Codes
Cognitive	Explicability	C1 - Intelligibility	right to explanation, transparency, opacity, black-box design, interpretability, explainability, comprehensibility, traceability, intelligibility, accuracy, efficiency, quality, reliability, minimise error, reduce risk, detecting causality than correlations, trust in an algorithm, faith, fidelity, generality, scalability,
	Explicability	C2 - Accountability	accountability, responsibility, liability, ownership of data and decisions, culpability
	Justice	C3 - Fairness	avoiding bias, fairness, justice, accessibility, discrimination, human rights, racial and gender stereotypes, information asymmetries, equality, freedom and justice, basic rights, equality, fair use, unfair outcomes
	Justice	C4 - Promoting prosperity	Socially beneficial, prudence, human values principle, common good, augment human capabilities than replacing them, the benefit of humanity, attention to context and culture
	Justice	C5 – Solidarity moral sensitivity, empathy, and appreciation for human rights	solidarity, empathy, social inequality issues, social justice
	Autonomy	C6 - Autonomy	autonomy, choice, human free agency, freedom of choice, nudging, power of user
Physical	Beneficence	P1 - Dignity and well being	temperance, responsible leadership, stakeholder rights, human dignity, how potential users are treated, loss of agency, compassion, genuine concern, datafication of society

	Beneficence	P2 – Safety	safety, mitigate harm, reduce fatality, damage
	Beneficence	P3 - Sustainability	a positive view of the future, respect for the public good, utilitarian, energy use, environmental effects, sustainability of the planet
Information	Non-maleficence	I1 - Privacy	privacy, access to personal data, surveillance, digital rights, trust, consent, datafication
	Non-maleficence	I2 - Security	security, data protection, confidentiality
Governance	Governance	G1 - Regulatory impact	avoid deception and coercion, policies to reduce social injustice, human rights and victim access to an effective remedy, intellectual property, data ownership, occupational rights, surveillance, consent
	Governance	G2 - Financial and economic impact	positive benefits in the marketplace, reduce cost, data monetisation, additional revenues, antitrust factors, corporate digital responsibility, national and international economic impacts
	Governance	G3 - Individual and societal impact	shifts in society with technological advancement, change in cultural and personal values, moral consequences, unemployment, retraining of displaced workers, deskilling, basic existential principles of humanity and society, social effects, the wealth gap, digital gap, job displacement and replacement, isolation, deprofessionalisation

Encompassed under the physical domain are the ethical implications pertaining to the introduction and evolution of new technologies. The eneficence principle is only discussed in this domain. Some of the ethical implications listed include well-being and sustainability. Under the cognitive domain, the ethical implications identified pertain to the use of algorithms. The principles of explicability and justice are relevant here, with their associated implications including solidarity, fairness and promoting prosperity. The information domain incorporates ethical implications surrounding privacy and security, which are core to the non-maleficence principle. Hence, non-maleficence is the only principle discussed under the information domain.

As the sixth ethical AI principle, governance is seen as “the practice of establishing and implementing policies, procedures, and standards for the proper development, use, and management of the infosphere” (Floridi 2018, p. 3). Governance incorporates both the formal and informal rules, alongside morals and values which are conveyed within the ramifications of regulatory, economic, and societal impacts.

Regulatory impacts pertain to relevant laws and legislation that regulate AI behaviour (Floridi 2018), the economic impact highlights the economic effects of AI diffusion on adopting organizations (Grewal et al. 2020), while the individual and social impact looks at the changes in individual agency due to the transformational impacts of AI on society (Akter et al. 2021). Examples of such impacts include relevant human rights regulations, revenues accrued and cost savings, and changes in society to address new technologies respectively.

Hence, the framework presented in Table 2 combines elements of the AI4People Framework by Floridi et al. (2018), alongside several ontological frameworks and ethical implications discussed in literature and introduced by Ashok et al. (2022). The addition of governance as the sixth principle allows better understanding of the interlinkages between the role of regulation and its impact on ethics. The Extended AI4People Framework is used as the main theoretical background for this research. In particular, the codes provided in the Framework are useful heuristics for analysing citizen perceptions of ethical implications of AI in the public sector, as citizens may use different terminologies to refer to similar concepts or focus on relevant implications.

The theorical background acts as a guiding lens to analyse how citizens experience the practical ethical implications of AI use, which principles these implications pertain to and the ontological domains they concern, which have great social impacts. Ashok et al. (2022) argue that this categorisation gives greater insight into the ethical implications affecting each domain, so that if negative outcomes or tensions arise, certain domains or ethical principles can be prioritised based on the context.

2.4 Shortcomings of Ethical AI Frameworks

The establishment of ethical AI frameworks follows the recognition that technology plays a crucial role in peoples' daily lives, with some of their outcomes regarded as neither inherently positive nor preferable (Héder 2020). Most crucially, the assumption that technology can be controlled serves as the basis for these frameworks, which contribute to setting the direction of what should be done from a moral perspective (Floridi 2018). Ethical frameworks thus aim to provide guidelines on the use of new technologies, which if left unattended, may have far-reaching economical and societal consequences, including value erosion (Héder 2020).

Consequently, Floridi et al. (2018) argue that establishing high-level principles is crucial to ensure that outcomes of AI are for society's benefit. These principles aid with summarising compound ethical issues into key statements that can be comprehended and accepted by diverse societal actors. Principles are used as a method to show formal

obligations to ethics by organizations, with the aim of addressing public concerns and promoting ethical commitments. This is done by following a soft-law approach, wherein the ethical principles are not regarded as legally-binding, but instead uphold a persuasive character (Heilinger 2022).

Despite this, scrutiny is still voiced by a growing number of researchers about the shift to ethical AI principles, with researchers deeming them ineffective (e.g., Lauer 2021). For instance, Mittelstadt (2019) argues that the high-level principles are too abstract to be action-guiding. His sentiment is echoed by Metzinger (2019), member of the EC HLEG AI, who stated that the "guidelines are lukewarm, short-sighted and deliberately vague", alongside "ignoring long-term risks, glossing over difficult problems" and "violating elementary principles of rationality" (n.p.).

Although principles play a valuable role, the statement highlights that they are not enough to ensure that societal benefits can be reaped, and the risks associated with the introduction of new technologies can be mitigated. Hence, two main criticisms pertaining to ethical AI principles are discussed in academic research: (1) they are too general and abstract to create a common understanding regarding their meaning and (2) a principle-implementation gap currently persists, leading to ethical tensions.

2.4.1 Operationalisation

Ethical AI principles are the result of initiatives that aim to express general societal values into broadly accepted, high-level principles that guide people and processes relating to AI implementation and governance in diverse contexts. These principles are at the intersection between values, ethics and technology, which together contribute to the advancement of a moral background to which elements regarding AI projects can be compared to (Hickok 2021). Accordingly, the value of the general, high-level principles lies in their attempt to identify relevant moral themes applicable in a multitude of contexts (Anderson and Anderson 2007).

Nevertheless, the ethical principles are incorporated into the systems by AI developers, who Munn (2022) highlights possess diverse backgrounds and come from different disciplines, leading to discrepancies in their incentives, histories and moral obligations. Hence, although there may be a consensus on which ethical AI principles should be upheld in practice, differences may emerge regarding "how the principles are interpreted; why they are deemed important; what issue, domain or actors they pertain to; and how they should be implemented" (Jobin et al. 2019, p. 396).

Here, the issue pertaining to the operationalisation of principles is two-fold. First, the ethical AI principles are conveyed using vague terms, rendering them ambiguous and hard to interpret, hindering the emergence of a common understanding regarding their definitions. For instance, beneficence is defined under the Montreal Principles (2017) as promoting the "well-being of all sentient creatures" (p. 545). In this regard, Munn (2022) argues that while the use of information and communication technologies supported human flourishing by enhancing communications and business ventures, they simultaneously caused harm to the environment through degradations, emissions and contributions to the climate crisis. In this context, AI can be seen as furthering the well-being of humans at the expense of other sentient beings.

Beneficence also denotes the importance of sustainability to ensure that AI promotes the essential conditions required to sustain life on Earth. However, different stakeholders hold different interpretations as to how this objective is realised. Whereas environmentalists encourage the use of AI to optimise energy, resources and stay within planetary boundaries (e.g., Steffen et al. 2015), neoliberal proponents argue that economic growth is what is required to sustain life, as shown by how oil conglomerates claim their work furthers sustainability (e.g., Desai et al. 2021). Hence, the ambiguity of the concept results in different definitions by actors that align with their activities and interests.

Second, the issue with operationalising principles lies in trying to translate normative concepts into technical rules and best practices, which is greatly underestimated within the technically-oriented industry. For instance, Hagendroff (2020) argues that technical fixes for commonly recurring ethical issues already exist, or can be easily developed. He elaborates further on how "accountability, explainability, privacy, justice, but also other values such as robustness or safety are most easily operationalised mathematically and thus tend to be implemented in terms of technical solutions" (Hagendroff 2020, p. 103). This technical view upholds a checklist approach to ethics, in the absence of meaningful reflexive practices on what technological implementations of high-level principles entail. Hickok (2021) argues that the main critique of this approach is that it promotes technical fixes to societal issues and does not consider the underlying social and economic situations surrounding these issues.

Translating high-level principles into norms and requirements to be applied in a specific context yields normative considerations that AI developers must decide on in the manner they deem appropriate, in the absence of coherent roadmap or implementation plan. Although a high-level consensus between actors regarding ethical AI principles is present, Mittelstadt (2019) highlights that these principles are too vague to give meaningful guidance. Diverse interpretations of the principles (e.g., beneficence) exist, which result

in significant differences between requirements in practice. Such discrepancies in requirements emerge when the principles are first translated and implemented in a concrete setting.

Clouser and Gert (1990) argue that such an approach to principles hides the presence and significance of moral disagreements found in society for the sake of developing a comprehensive theory. This leads the ethical AI principles towards a position of moral relativism. Moral philosopher McIntyre (1988) put forward two main questions when exploring the role of moral reasoning when constructing moral machines: *whose justice?* and *which rationality?* Modifying these for the AI context leads to questions regarding *whose morality?* and *which rationality?* do ethical AI principles uphold (Serafimova 2020).

To (partly) overcome this issue, de Bruin and Floridi (2017) suggest that AI developers should be aware of the required technical details within their frameworks, which requires reflections on how the data is created, handled, shared and utilised. Kitchin (2017) also emphasises these reflective practices when coding and designing algorithms. Hagendroff (2020) argues that for these steps to be followed sufficiently, a (partial) shift by AI developers to *micro ethics* should occur. This shift entails that changes to the abstraction level of the principles ought to occur if ethics are to fulfil their aim of achieving their desired influence and impact in society (Morley et al. 2020). This is relevant as clearly articulated principles are more beneficial for guiding practitioners and promoting self-governance within developers (Mittelstadt 2019).

Nevertheless, remedying issues pertaining to the vagueness of the principles and translating normative concepts is made more difficult due to the *ethics shopping* approach suggested by Floridi et al. (2018). Ethics shopping entails that both public and private actors borrow terms from various disciplines and fields, which furthers the presence of fuzzy principles through a mix-and-match approach that best matches the actors' interests and rationalise their behaviour. This is done instead of scrutinising actors' behaviours in reference to publicly upheld ethical standards. Consequently, the adoption of ethical AI principles can be regarded as no more than just virtue-signalling, used to defer laws and regulations while shifting debates to vague issues, abstract principles and technical solutions (Greene et al. 2019).

2.4.2 Principles-Implementation Gap

According to Morley et al. (2020), several challenges persist when ethical AI principles are implemented in practice as a result of their subjectivity, complexity, and lack of standardisation within the definition of each ethical principle. Schiff et al. (2020)

summarise five main concerns regarding why AI principles are difficult to implement and do not yield effective practices.

First, AI has complex impacts on society that are underestimated. AI developers often focus their efforts on a single product and its associated harms, without considering the wider societal and economic consequences. Second, dubbed as the *many hands problem* by Floridi (2013), the large number of stakeholders muddles where the distribution of accountability lies. Whereas AI developers regard their role as ensuring the quality of products, businesses regard their role as accruing revenue, hence creating a gap on who tackles key societal impacts of AI. Third, the diversity of stakeholders from different disciplines and fields yields a variety of frames, attitudes and values. Hence, the resultant perspectives are regarded as either being too narrow or wide, or promote solutions that prove challenging to translate.

Fourth, the over-abundance of methodologies and tools present for the ethical use of AI poses difficulties for evaluating each's utility, or easily comparing it to other established methodologies or tools. Mittelstadt (2019) even argues that tools that are yet to exist cannot be tested effectively to ascertain which works best in what context. Fifth, the functional separation between the technical teams (i.e., AI developers) and non-technical teams (i.e., compliance staff) builds obstacles. Whereas the former may not have adequate insight on how normative considerations are weighed to incorporate these concepts into systems, the latter may struggle with comprehending or modifying the designs of AI systems alone.

Whittlestone et al. (2019) argue that the gap between the ethical principles and their implementation is exacerbated when considering that the principles will inevitably lead to the rise of ethical tensions. The authors define tension as "any conflict whether apparent, contingent or fundamental, between important values or goals, where it appears necessary to give up one in order to realise the other" (p. 197). Tensions arising as a result of the implementation of principles may present a moral trade-off, wherein two goals or values cannot be pursued simultaneously.

For instance, the need for AI systems to be trained with large, diverse datasets to minimise the emergence of biased results may impede on individuals' need for autonomy and privacy, as their data is increasingly used and remain out of their control (Hagendroff 2020). Similarly, the utilisation of risk-benefit evaluations may result in diverging outcomes based on considerations of whose well-being is pursued, alongside by which stakeholder and their interests. These examples highlight the difficulty of implementing ethical AI principles in practice, which heighten uncertainties regarding the prioritisation

of principles and approaches to resolving conflicts. Jobin et al. (2019) indicate that such difficulties may undermine efforts of establishing a global ethical AI agenda.

Beauchamp and Childress (2001) propose that for principles to be action-guiding, they need to be supplanted by an explanation on their application in distinct scenarios, alongside how they are balanced when tensions emerge. This entails that the implementation of principles needs to consider the social and cultural context (Greene et al. 2019), established structures and practices (Latour 2007), alongside adopting a tension-focused approach (Whittlestone et al. 2019). Such an approach aids in emphasising which ambiguities exist between principles and which gaps in knowledge regarding the societal impact of AI prevail, in the absence of a universally established hierarchy between principles. Due to the fact that similar tensions are expected to emerge across a variety of cases, Whittlestone et al. (2019) argue that a tension-focused approach is beneficial as it is not case-specific, nor overly reliant on vague high-level principles. Accordingly, the authors point out four main tensions apparent in practice:

> **Tension One:** *Advancing the efficiency and quality of services versus safeguarding citizens' privacy.* This tension relates to the large number of datasets required for AI to work sufficiently, which comes at the expense of needing to collect and use citizens' data.
>
> **Tension Two:** *Accurate predictions for individuals versus fair treatment.* This tension relates to the fact that AI can improve services delivered to citizens based on future predictions of what their lives will look like, but the predictions may be biased or not established for sub-groups where representative data is scarce.
>
> **Tension Three:** *Personalisation versus citizenship.* This tension relates to data collected about individuals which can help provide more personalised services, but at the cost of differentiating between citizens, threatening citizenship ideals.
>
> **Tension Four:** *Benefits of automation versus safeguarding dignity.* This tension relates to the fact that automation reduces the amount of mundane tasks that are done by administrators, but at the risk of disrupting established practices, homogenisation and deskilling of society.

Whittlestone et al. (2019) also underscore the importance of realising when tensions occur, then deploying necessary actions and efforts to mitigate, or at least minimise their negative effects. Although not all tensions are resolvable, trade-offs will be needed to prioritise certain values over others. This can be perceived as a crucial step to bridging the gap between high-level principles and their implementation in practice. Nevertheless,

prioritising principles and values is inherently a political process, and should incorporate the voices of different stakeholders and members in society.

The criticisms towards the ethical AI principles highlight that while a global consensus is sought within the ethics domain, this should not come at the expense of ensuring moral and cultural pluralism relative to specific contexts (König 2022). In this regard, the principles ought not be perceived as comprehensive checklists that ensure ethical decision-making, but instead as general ethical considerations that are tailored to diverse societies' specific needs. This may entail the emergence of tensions, which reiterates the importance of prioritising values based on differing needs. Such a perspective may reduce costly ethical AI mistakes, heighten public acceptance in the systems and provide benefits to society. As Hagendroff (2020) states this:

> A transition is required from a more deontologically oriented, action-restricting ethic based on universal abidance of principles and rules, to a situation-sensitive ethical approach based on virtues and personality dispositions, knowledge expansions, responsible autonomy and freedom of action (p. 114).

2.5 Role of Citizens

The critiques regarding ethical frameworks highlight that the field of AI ethics consists of various value tensions and moral trade-offs. Often, it is up to the developers of AI systems to form, execute and voice decisions in the absence of a coherent implementation plan (Hickok 2021). As such, when analysing governance mechanisms relating to AI, Cath (2018) poses various important questions such as "who sets the agenda for AI governance? What cultural logic is represented by that agenda? And who benefits from it?" (p. 4). These questions are relevant considering Hickok (2021) argues that ethical AI principles have thus far failed to include the opinions of people targeted by AI systems.

Powles (2018) posits that citizens have a vested interest in the responses to these questions. The EC (2019b) puts forward the overarching objective of AI as "in the service of humanity and the common good, with the goal of improving human welfare and freedom" (p. 4). Kieslich et al. (2022) argue that the pursuit towards a human-centric AI is conveyed in this objective, which entails that the perceptions of people targeted by AI systems should be incorporated. Nevertheless, the lack of inclusion of relevant societal members leads to ambiguity regarding which priorities and values are pursued. A power asymmetry is present between developers and people impacted by their decisions (Whittlestone et al. 2019). In this regard, Coeckelbergh (2020) suggests that "if we endorse the ideal of democracy and if that concept includes inclusiveness and

participation in decision-making about the future of our societies, then hearing the voice of stakeholders is not optional but ethically and politically required" (p. 170).

Accordingly, Habermas's (1991) argument of discourse ethics may tackle issues relating to inclusion and pluralism, as shown by Buhmann et al. (2019) in their research regarding the management of algorithmic accountability. Habermas (1991) argues that societal norms and morals are not hierarchically established, but instead are the result of discussions between societal members who present different views, consider each other's perspectives, justify and reassess their positions to reach a universally agreed upon decision. This perspective is relevant nowadays due to rising interest in the social acceptance of new technologies (e.g., Taebi 2017), in the aftermath of public opposition and controversies.

For various decision-makers, public opposition is regarded as a mere obstacle to the implementation of new technologies that must be overcome. Hence, altering public opinion is pursued through "[using] marketing methods [to] maximise the likelihood of [technology's] successful introduction" (Schulte et al. 2004, p. 677). Such methods ignore the root of the issue – understanding the factors influencing the perceptions of new technologies. In this regard, Taebi (2017) argues that both ethics acceptability and social acceptance are important elements to consider. The former is defined as "reflections on new technology that take into account the moral issues that emerge from its introduction" while the latter considers the acceptance (or at least, toleration) of a new technology by a community (p. 1818). Both these concepts are interlinked, as before a technology is accepted in society, the ethical considerations pertaining to it are discussed and reflected upon.

Accordingly, a growing number of research suggests relevant criteria for assessing ethics acceptability (e.g., Asveld and Roeser 2009), which impacts social acceptance. Taebi (2017) puts forward three main criteria: right to consent, integration of stakeholders' opinions to ensure pluralism and their contextual knowledge. The first criteria refers to citizens having the moral right to be informed about potential risks and have the ability to consent to such risks. The right to be informed is established in the Aarhus Convention, which dictates that access to information and public participation in decision-making is upheld (Cramer 2008). Nevertheless, it may not be realistic for every individual impacted by AI to provide their consent, as society cannot function if every possibility of risk is removed (Asveld and Roeser 2009). Despite this, Taebi (2017) argues that pluralism should at least be acknowledged and considered when making decisions.

The second criteria advances the need for pluralism, a key democratic element, to highlight the various moral opinions present. Within the context of AI, pluralism refers

to the various moral and cultural values in a society. The final criteria argues that stakeholders can have crucial local knowledge and information, emphasising the need for their opinions to be considered. Unlike widespread assumptions that citizens' understandings of risk are illogical and emotional, Asveld and Roeser (2009) argue that citizens can provide invaluable insight into issues due to their exposure to risk-related ethical outcomes.

The presence of value tensions and moral trade-offs highlights the need to address the power asymmetry between AI developers and people impacted by the outcomes of these systems. Ethics acceptability entails that citizens can reflect on the ethical and moral aspects regarding the introduction of AI, which is considered a crucial element by various ethical guidelines and researchers. Incorporating citizens' opinions regarding AI upholds pluralism within a society, which is a key democratic element. Once citizens adequately reflect and voice their ethical concerns, this presents an opportunity to make changes that enhance citizens' trust in new technologies. This may aid motivate the social acceptance of AI as it is increasingly implemented in public sectors worldwide, allowing for more of its benefits to be reaped.

In conclusion, although research regarding AI is not a new effort, it only recently gained momentum in the public sector. While the use of AI promises efficiency and cost benefits, it also requires value-based judgements and ethical considerations. These impact citizens' perceptions and judgements regarding AI projects. Value-based and ethical elements are often discussed within ethical guidelines and principles. Although various are put forward, the five ethical principles by Floridi et al. (2018), and their further categorisation by Ashok et al. (2022) are highlighted and serve as the main theoretical framework of this research to which citizens' answers can be compared to.

Critiques against ethical approaches are also discussed, as they serve to shed light on the shortcomings of current practices. These elements highlight the contentiousness relating to ethics and AI, emphasising the importance of incorporating citizens' views on the matter as individuals impacted by AI outcomes. Citizens' perceptions are elaborated on in the coming sections, in relation to the Extended AI4People Framework and the four main ethical tensions.

3 Research Design

The ethical implications relating to the use of AI in the public sector are studied within the context of the Netherlands. The reason the Netherlands serves as an interesting case study are three-fold: (1) focus on digitalisation, (2) empowerment of municipalities, and (3) established importance of ethics. These three aspects shaped the way that AI is implemented in and perceived within the Dutch society. By elaborating on these, a more nuanced understanding on how the Netherlands reached its current AI position and ethics in the public sector can be discerned. Flyvbjerg (2006) argues that underpinning salient elements can help create a narrative, which aids with understanding experiences.

3.1 Case Study

In terms of digitalisation, the Netherlands continues to promote a culture of innovation and actively prioritises this at the national level. These elements are underpinned in the Dutch Digital Strategy report, which highlights governmental digital policies and is updated on a yearly basis (Rijksoverheid 2021). As a result, the 2022 Digital Economy and Society Index (DESI) shows the Netherlands ranking third in the European Union (EU), which is one spot higher than the country's ranking for the past two years. Particular areas where the Netherlands excels include human capital, connectivity and the integration of digital technology, such as in the provision of digital public services. The DESI (2022) report shows positive perceptions of the digital advancements of the Netherlands and anticipates positive spill overs across EU countries, as shown in the following statement:

> As the fifth-largest economy in the EU and top performer in DESI rankings, the Netherlands' progress in digital transition over the coming years will be crucial to enable the EU as a whole to reach the 2030 Digital Decade targets (p. 4).

Digitalisation efforts are not delimited to the national scale. As a result of increased decentralisation between 2014-2015, local municipalities' responsibilities and tasks were expanded (Franzke et al. 2021). This was done to increase government efficiency by delegating social responsibilities to the local level, altering the relationship between municipalities and their residents through heightening the former's influence. Municipalities thus felt incentivised to innovate their processes in order to tackle the increased workload (Maarse and Jeurissen 2016).

Municipalities also emphasised that citizens' data should be handled in an ethical and responsible way (van Noort 2015). This resulted in many initiatives, including in the

municipalities of Amsterdam, Utrecht and Zaanstad. For example, the Amsterdam Municipality published an ethical data manifesto, made publicly available a register of algorithms used and issued guidelines relating to algorithm procurement (Franzke et al. 2021). This shows the ethical turn on the local level, established for the past eight years.

Despite the focus on ethical data usage at the local level, widespread critiques are posed against the Dutch government at the national level. These critiques followed reports of the Dutch childcare benefits scandal (*toeselganaffaire*) in the media, which still serves as a cautionary tale arguing for the establishment of ethical guidelines and reiterating ethics importance (Amnesty 2021). The scandal brought to light how the government-developed algorithm used to detect social benefit fraud, deployed since 2013, discriminated against ethnic minorities and individuals with immigrant backgrounds by falsely labelling them as fraudsters. The algorithm was later outlawed by a Dutch court, as it was considered discriminatory and violating fundamental human rights (Brown 2020).

This scandal contributed to the decreasing public trust in the national government. In the aftermath of the social destruction caused by the algorithm, the Dutch State Secretary of the Interior and Kingdom Relations (2019) published a policy brief on the use of AI, which was presented to the Dutch Lower House. The brief emphasised the recognition of relevant opportunities and risks relating to AI, while safeguarding public values that are based on fundamental human rights. Special attention is paid to AI applications that have an impact on individuals and society. The public values highlighted are non-discrimination, privacy, freedom of expression, human dignity and personal autonomy.

The focus on digitalisation and upholding ethical guidelines at both local and national levels aided in shaping the current environment of AI use in the Dutch public sector. Currently, the establishment of a National Growth Fund Investment Programme expedited the accruement of €2.1 billion investments in AI for 2021–2027 (Rathenau 2021). This contributed to the Netherlands ranking 8th in the 2022 Global AI Index and 1st amongst EU countries (Tortoise Media 2022). The Index captures the capacity of AI in the analysed countries through three main pillars of analysis: implementation, innovation and investment.

The Index also highlights that the Netherlands is not only aiming to increase AI implementations, but also actively attempts to weave the novel technology into the societal fabric. In this regard, considering the previous ethical breach and attempts by the Netherlands to embed AI into society whilst ensuring fundamental rights are upheld, citizens can play a greater role in setting the direction of future AI implementations. Particularly regarding the ethical implications of AI, citizens are crucially affected by these and have a greater stake in ensuring their rights are upheld. This view was

emphasised in a Deloitte report about AI ethics, which stated that “AI systems’ behaviour should reflect societal values, [thus] gaining societal consensus on the ethics of AI is one of the key tasks of the government” (Hashimi 2019, p. 8).

As such, the incorporation of citizens’ opinions sheds light on how they perceive the current AI climate in the Dutch public sector. This provides insight into what public values citizens prioritise, alongside their perspectives on the benefits and risks of current AI applications. AI developers and practitioners can use these insights to designate potential areas for improvement in implemented and future AI projects, with hopes of advancing the social acceptance of AI in society.

3.2 Focus Group Method

Focus groups are used as the main method to illicit responses from Dutch citizens regarding their perceptions of the ethical implications of AI use in the public sector. This method is defined as “an informal discussion among selected individuals about specific topics” (Beck et al. 1986, p. 73). Though minimal, the definition posits an interactive element as participants engage with each other. The moderator does not play an active role in the discussions as this heightens risks of embodying preconceived ideas and biasing answers. Instead, the moderator is provided the opportunity to discern how participants construct, express, defend and modify their views throughout, alongside observing how the process of collective sense-making occurs (Wilkinson 1998).

These elements show that the emphasis in focus groups is placed on the participants rather than the interviewer (i.e., moderator), who are with provided a platform to share their (divergent) perspectives and views (Guba and Lincoln 1994). To enhance the utilisation of focus groups as a main method, Morgan et al. (1998) discuss four main steps that need to be undertaken: (1) design, (2) data collection, (3) analysis, and (4) results and reporting. Figure 1 shows an overview of these steps, alongside the considerations each step entails.

3.2.1 Design

The process starts by defining the main research objectives, which sets the direction for the design of the focus group. As the purpose of the focus group is to explore insights on citizens’ understandings and views on the ethical implications of AI use in the Dutch public sector, a phenomenological approach is undertaken. Wilkinson (1998) argues that this approach is relevant for research which is primarily concerned with peoples’ own perspectives and understandings in relation to the issue(s) under focus. In this regard, pre-existing conceptions and notions of the moderator do not easily materialise due to the fact that collective sense-making occurs in the interactions between participants.

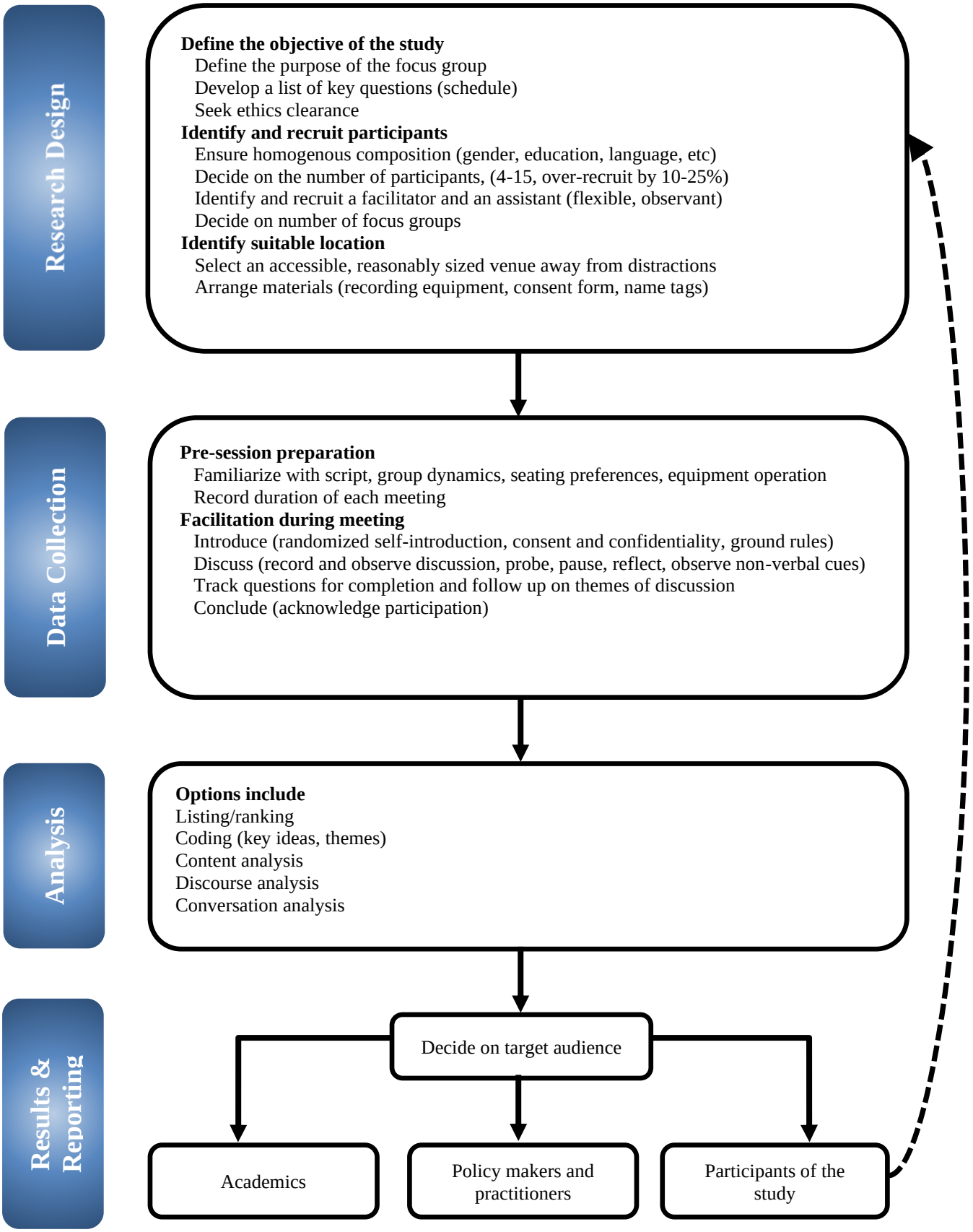

Figure 1 – Focus Group Method Steps by Morgan et al. (1998)

While the phenomenological approach is the primary approach to the focus groups, the discussions also aim to empower participants. The participatory action approach towards focus groups is especially relevant for gaining insight into the views of people who are not adequately represented in research or whose voices are otherwise ignored (Plaut et al. 1993). Johnson (1996) highlights that focus groups conducted within this approach serve "to empower and to foster social change" (p. 536), resulting in the emergence of a collective understanding, and usually solutions to these societal problems.

A list of questions (i.e., schedule) to help structure the focus group discussions is then developed following the identification of the research objectives. Krueger (1988) recommends that the schedule has no more than ten questions, and usually contains around five to six questions. The Appendix provides an overview of the schedule used in this research, which contains a total of eight questions. The schedule was developed according to Krueger's (1988) guide on the categorisation of questions to pose, which follows a funnel approach from highly general to more specific questions. Accordingly, five categories are presented: opening questions, introductory questions, transition questions, key questions and ending question.

Opening questions are used to gauge participants' knowledge of the subject, and sets the tone for the rest of the discussion. The questions are meant to be highly general as to encourage discussions and provide an opportunity for the participants to open up. An example of an opening question posed in the focus groups conducted is *can you tell me what comes to mind when you hear the term artificial intelligence?* and *what do you know about how AI is used in the public sector in the Netherlands?*

Before proceeding to the following category of questions, some time was taken for the moderator to provide participants with additional information regarding AI uses in the Dutch public sector, as suggested by Breen (2006). Two AI implementation cases are briefly presented, which helped participants with orienting their answers along the discussion. The cases highlight both negative (i.e., *toeslagenaffaire*) and positive (i.e., notifications public space Amsterdam; SIA) implications of AI use, and were selected on the basis of two main criteria. First, both are relatively recent and covered in the media since 2018. This suggests that citizens were subjected to these cases beforehand and are more likely to hold opinions on the matter. Second, providing both a positive and negative example reduces the risk of swaying opinions in one direction and allows participants to contrast case elements to each other. This can provide fruitful debates on the topic and aid in discerning elements that citizens prioritise and find problematic.

Following elaborations on the cases, the remaining questions are introduced, starting with more general questions then honing in further as the discussion continues. Participants

are asked *what are important things to consider for the public sector when using AI?* as an introductory question to the topic. Questions similar to these encourage initial dialogue and brainstorming between participants about important elements, which are often built upon and recur throughout the discussion. These initial dialogues later aid in the emergence of more concrete views and opinions as the discussion progresses, which is key to obtaining relevant answers to key questions later posed.

Based on the established main and sub-questions of this research, key questions pertained to gaining insight on participants' understandings of ethical principles (e.g., *how important are ethical aspects of AI?*) and priorities pursued when ethical tensions emerge (e.g., *what is the one most important thing for the public sector to prioritise when using AI?*). Following this, to finalise the discussion, participants are presented the opportunity to share any final thoughts as an ending question before the focus group is adjourned. As the topic is not considered sensitive, nor the questions introduced or the answers collected from the participants, no ethics clearance is required.

Identification of participants is the next step, which Kitzinger (1994) highlights is critical for determining group dynamics and impacts the data generated throughout the focus group. Krueger (1988) also emphasises the importance of participants sharing similar characteristics since homogenous compositions are regarded as positively effecting group dynamics. Due to the main aim of this research, education and age range were the two characteristics most emphasised when identifying participants. The former to increase the probability that participants have relevant knowledge about the complex topic of ethical implications of AI, and the latter for participants to relate to each other and have similar experiences to share.

The identification of participants is followed by their recruitment, which can be done in various ways. Convenience sampling is used in this research to recruit participants who are accessible, geographically close and showed willingness to discuss the topic. These criteria are generally accepted for when using this sampling method (Krueger 1988). Although this does not yield a randomised representative sample of the population, convenience sampling is relevant in this research due to the aim of exploring initial perceptions of citizens, instead of producing generalisable results for the Netherlands.

Three focus groups were conducted with Dutch citizens to gain insight about their perceptions of the ethical implications of AI use in the public sector, with this number in line with research recommendations (Burrows and Kendall 1997). Although five groups were initially planned, theoretical saturation was reached within the participants' responses wherein no new information was produced (Guba and Lincoln 1994). There were thirteen participants in total; five in the first group and four in subsequent ones.

Table 3 provides demographic data for the participants. Although between six to eight participants per group is recommended, smaller group sizes are favoured when the topic is complex as to allow adequate room for in-depth discussions (Krueger 1988).

Table 3 - Demographic Information

Gender	**#**	**%**	**Education**	**#**	**%**
Male	8	61.5	High School Graduate	1	7.7
Female	5	38.5	Bachelor's Degree	3	23.1
Age	**#**	**%**	Master's Degree	9	69.2
20 - 35	5	38.4			
35 – 55	2	15.4			
>55	6	46.2			

An assistant accompanied the moderator throughout each of the focus groups, with the primary responsibility of taking down notes on the participants, their body language, non-verbal cues, general discussions, and ensuring that the time limit for the focus group is upheld. On average, each focus group lasted for one hour and ten minutes, which is in line with expectations communicated to the participants in advance. To ensure that participants remained comfortable and focused during the discussions, the focus groups were held in the moderator's house. This is in line with recommendations that the location is accessible, comfortable, reduces distractions and has enough seating so that participants can visibly see each other (e.g., Smith 1972). Participants were provided the opportunity to consent before the focus group was initiated. After obtaining consent, a mobile device was placed centrally between the participants to record the discussion.

3.2.2 Data collection

According to Burrows and Kendall (1997), a focus group requires to be led by both a moderator and assistant. The former is key for facilitating interactions between participants in a comfortable environment. The latter is relevant for taking notes, focusing on elements such as non-verbal cues, group dynamics and interesting findings. These notes supplement the data, and thus offer thicker descriptions about the focus group discussions in comparison to just examining verbal data (Fonteyn et al. 2008).

In each group, the participants were previously familiar with each other and thus required no introductions. Due to their relationships, the participants felt comfortable to share information and were relaxed throughout. Prior to the discussion, participants were reminded that their confidentiality will be upheld, and their participation is voluntary. Thereafter, the research topic was discussed, and the moderator only intervened when following up on emergent themes. The discussion lasted on average around one hour and ten minutes. Participants were provided the opportunity to make final comments towards the end, alongside being acknowledged for their contribution.

No major changes relative to the manner each focus group was conducted in occurred. While conducting the first focus group, special attention was paid to the wordings of the questions, how participants understood and responded to the questions posed, alongside additional practical elements such as time allocation. In doing so, relevant changes that enhance participants' experiences can be identified and established, as proposed by Krueger (1988). A minor change was introduced in the subsequent focus groups, which was in relation to honouring the time limit communicated. Following the question *how important are ethical aspects of AI?*, every participant was within a set order invited to summarise their key thoughts on the ethical aspects of AI in a few sentences. This is unlike the first focus group, wherein participants freely shared their opinions, resulting in new discussions that were off-topic and lengthy. Adopting this change resulted in responses from participants in a more orderly manner, while staying true to the question on hand.

3.2.3 Analysis

An interpretative phenomenological approach (IPA) was used when analysing the transcripts of the focus groups. IPA is relevant when examining individuals' lived experiences, alongside the meaning behind and sense-making of these experiences (Smith et al. 2009). Reid et al. (2005) suggest that the small group sizes and homogenous sampling are conducive towards IPA as to allow participants meaningful opportunities to discuss the topic in-depth. In addition to this, the use of IPA indicates that convergences and divergences between participants' answers in each group can be compared, highlighting recurring themes and the ways in which these themes emerge.

Smith et al. (2009) describe IPA as incorporating a *double hermeneutic process*, wherein participants make sense of their experiences, and the researcher attempts to make sense of their responses. Regarding the latter, although an inductive approach is commonly used with IPA, a deductive approach is instead adopted in this research. Smith et al. (2009) recognise the possibility of utilising a deductive approach, though this should be secondary to understanding participants' lived experiences in their words. Additionally, Byrne (2022) argues that in pursuing deductive approach, "the relationship between different items in the data set to identify recurring commonalities with regard to a conceptual framework" can be discerned (p. 1397).

A thematic analysis of the data through a deductive approach is used in this research, which assumes a top-down application of pre-determined codes on the data. This allows the analysis to be aligned with the goal of preserving focus on and answering the research questions through the systematic application of the theoretical framework (Bingham and Witkowsky 2022). The pre-determined codes are obtained from two sources: the codes in

the ExtendedAI4People Framework presented in Table 2 and the ethical tensions put forward by Whittlestone et al. (2019). These codes are incorporated into a codebook, which includes a list of the nine main codes used (i.e., five principles with each's sub-codes and four tensions), in addition to a detailed description of these codes, including what is included and restricted.

The codebook approach to thematic analysis results in the development of themes prior to engaging with the data (Byrne 2022), thus enhancing alignment with the research questions. This allows the theoretical framework to act as a guiding lens to better understanding citizens' perceptions of the ethical implications of AI, and aids in centring the analysis around established ethical AI principles rather than imposing meaning on the data.

The analysis of the transcripts was done in line with Braun and Clarke's (2006) six-phase framework for undergoing thematic analysis: (1) become familiar with the data, (2) generate initial codes, (3) search for themes, (4) review themes, (5) define themes, and (6) write-up. Although these phases portray a logical sequence, data analysis may not follow a linear order. This allows for iterative analysis, wherein movement between the phases is done when necessary. Hence, the six phases ought not be perceived as hard rules, but instead as general guidelines that can be iteratively applied to be aligned with the research questions and goal.

For the first phase, familiarisation with the data entailed reading and re-reading the focus group transcripts to get a first impression over what was said. In doing so, initial interesting information and passages can be identified that are relevant for addressing the research questions. Active listening is encouraged during this phase. Following this, codes are used to categorise the data, which were pre-determined based on the ethical principles and tensions highlighted. Codes serve the purpose of summarising key information into a few words, which is useful for examining relationships and patterns that address the research topic and reflect the variety of perspectives (Braun and Clarke 2012). The qualitative data analysis software Atlas.ti was used to code the data, which is recommended for research with thematic analysis (e.g., Friese 2019). The transcripts were subjected to three iterations of coding to ensure all relevant data was highlighted.

Themes are then searched for after the relevant data has been coded. This entails a shift from analysis of single data items to finding aggregate meaning across the entire transcript. Braun and Clarke (2019) advance the understanding that themes typically incorporate *domain summaries*, which are defined as "summaries of what participants said in relation to a particular topic" (p. 5). Hence, this phase encompasses understanding and actively forming relationships between codes, which in turn communicates

meaningful information relevant to the topic. Nevertheless, due to the deductive approach adopted in this research, this and the following step are not as relevant considering that the themes were pre-determined.

Hence, despite the deductive identification of the themes, they are still subjected to a recursive review. This is done through Patton's (1990) *dual criteria for judging categories*, wherein at one level internal homogeneity ensures that codes are aligned with the theme, and at the external level heterogeneity ensures that themes are relevant for answering the research questions. In total, three themes are established, and inform subsequent analysis. The themes are then defined and joined together to create a narrative, alongside finalising their naming to be short and descriptive.

3.2.4 Reporting

The final phase for thematic analysis (i.e., write-up) is similar to the reporting step of focus group discussions, thus both are elaborated upon together. This step occurs following data analysis, as the results of the previous phases are consolidated to produce a coherent report. Nevertheless, it should be noted that report writing is a recursive process, subjected to change as codes and themes evolve throughout the analysis. Key quotes put forward by participants in their own words regarding relevant points are highlighted and presented under three main themes: (1) general perceptions of AI ethical implications, (2) understanding and prioritisation of ethical principles, and (3) ethical tensions. This research is shared with the participants at the end to validate the results and increase its credibility, through the process of *member checking* (Guba and Lincoln 1994).

As such, the research design outlines relevant elements pertaining to the Dutch case study as to motivate its use and to address the research question regarding citizens' perspectives of the ethical implications of AI use in the public sector. This topic is considered novel, as citizens' perspectives are inadequately encompassed in current research. Accordingly, the emphasis on the phenomenological and IPA approaches within the analysis of data aids in understanding local perceptions in citizens' own words. Coupled with the use of a deductive approach to thematic analysis, this allows insights to be compared to existing literature, hence contributing to theoretical knowledge, alongside highlighting relevant areas where research is yet to explore. The approach also allows for practical recommendations to regarding the local Dutch context.

4 Results

The use of AI in the Dutch public sector is not a recent phenomenon. To date, there are various applications of AI both within the national government and at the local level, predominantly with the purpose of reducing public costs and enhancing the delivery of public services (e.g., social benefits). Despite this, a common perception in the Netherlands is that citizens are wholly unaware of concrete applications of AI in the public sector. This lack of awareness impacts how citizens perceive the use of AI, alongside its repercussions on their daily lives. Responses from the participants should thus be considered within this backdrop and are discussed under three main themes: (1) general perceptions of AI ethical implications, (2) understanding and prioritisation of ethical principles, and (3) ethical tensions.

4.1 General Perceptions of AI Ethical Implications

This theme discusses Dutch citizens' general perceptions of the ethical implications of AI use in the public sector. Their perceptions are shaped by their understanding of AI, alongside media coverage of use cases and implications. With the former, most participants cited that their understanding of AI relates to "robots". Whereas regarding the latter, participants reported that the topic of AI was too difficult to grasp, mainly due to convoluted reporting and lack of transparency. This is shown by the answer of one participant who proclaimed that "it is too complicated for me as a normal citizen to fully grasp [AI]". Another participant had lesser knowledge about the topic, claiming that "it is the first time I have heard them using AI in the Dutch public sector". As such, though lacking a solid basis for what the term entails and how AI is used, participants were still eager to discuss various elements that they deem important and should be safeguarded when the public sector uses AI.

Overall, participants reported mixed feelings about the use of AI in the public sector due to the resultant emerging ethical implications in society. Relevant snippets pertaining to this claim are participants sharing that AI is "scary, and hopeful" and that it is "something mysterious". Whereas the participants do believe that the use of AI has undeniable benefits both for the public sector and society in terms of efficiency and accuracy, questions and concerns persist regarding ethical elements. Accordingly, both the positive and negative perceptions are discussed under the following two sub-themes.

4.1.1 Positive Perceptions

Participants advanced three main positive aspects pertaining to AI use, which they regard as benefitting perceptions and understandings of AI use in the Dutch public sector. The

most cited positive use of AI amongst participants regards its ability to help make (more accurate) predictions that augment human knowledge and capabilities. Relevant areas mentioned include healthcare, mobility and infrastructure, as well as sustainability.

Within healthcare, it was repeatedly noted that "a lot of progress is being made, and a lot of benefits" exist within this domain. For instance, AI can be used to judge scans, better and faster detection cancerous cells, as well as reducing costs associated with "the whole health care system [which] is virtually collapsing on the demand globally for elderly, [as] it is not affordable or sustainable anymore". In this regard, AI serves the function of gathering medical complaints and aids with diagnosing, reducing pressures on hospitals and staff. Positive benefits are also reported within mobility and infrastructure, alongside sustainability. With the former, AI proves useful when providing inputs based on data points that improve "flows of traffic" and "to reduce accidents". With the latter, participants reported that AI can be used to produce climate models and "make more accurate estimations regarding the environment", which lessen the need for intensive human work.

Moreover, AI is also emphasised as enhancing the provision of services. This is done through improving collaborations between different PSOs, alongside increasing their efficiency. Currently, participants view the different organizations as encompassing data siloes, diminishing the flow of information. To this end, one participant noted that:

> Different parts of government have a lot of data that could be useful to other parts of the government, but are very closed off to that only that specific department… And if that would open up, then that would radically improve the government.

This improvement can already be seen, for instance by the tax authorities wherein most relevant information regarding citizens are pre-filled. Efficiency gains can also be the result of using AI in PSOs. Multiple participants noted these gains in the form of decreasing the intensity and time required for workloads since "going through those bureaucratic steps otherwise would be a very slow process". Particularly in the context of a labour shortage, AI can be conducive to increasing the productivity of existing people and automating routine jobs. This entails that citizens are able to automatically process simple requests online, rather than "having to queue or to interact with someone at the office just to get a document".

Though less common, the use of AI to enhance social elements is also reported. One participant noted how AI may aid in tackling one of society's wicked issues – loneliness.

This issue is exacerbated within the elderly population, who often feel isolated from the rest of society. The participant emphasised the following:

> I mean, if you look in elderly homes, everybody is lonely and all the people are alone, you know. In Japan and China, they put these robots there and people feel all of the sudden much more happy because there is a robot, maybe a furry robot, that talks with them and interacts with them.

Thus, these robots may utilise AI to personalise conversations with recipients and help tackle the aforementioned wicked challenge. In addition to this, one participant emphasised that contrary to more widespread beliefs, AI can be used to advance social justice issues. These can take various forms. For instance, one participant expressed how a PSO-deployed algorithm aided in improving their circumstances through providing evidence of misconduct by another individual, which otherwise would have been hard to obtain. In this regard, the participant claimed:

> I think it is a very helpful way sometimes and in some cases to help and give voice to those who might not be represented and might not able to reach out to anyone for their living circumstances or any other problem … in case you are the victim. It gives you an opportunity to at least show a problem you are facing first on a local level, then it could be expanded.

Similarly, AI can be deployed to bring more honesty when it comes to politics, and transparency when it comes to pursuing justice. With the former, one participant suggested that AI can provide decision support, and thus "telling more the truth to governments and policy makers". These may aid in the production of long-term oriented decisions in a more transparent way. Regarding the latter, various participants proclaimed that the use of AI when pursuing justice may be beneficial as judgments can be monitored and recorded. This is perceived as a better alternative to human judgement as "human racism, human errors and human misjudgement are still a greater problem than AI-led misjustice".

4.1.2 Negative Perceptions

Despite participants reporting on various beneficial uses and positive perceptions of AI in the Dutch public sector, negative perceptions dominate discussions, particularly regarding ethical concerns and implications of AI both at the personal level and on greater society. Participants initially expressed a lack of transparency by the government on concrete applications of AI in the public sector, which leads to a lack of awareness amongst citizens. Even when such instances are reported, the language used is often "so

complicated" that no common understanding between citizens emerges. Complicated language use is not delimited to reporting on AI use cases, but also related to aspects of AI and privacy. Participants alluded to the fact that the Dutch government collects excessive amounts of personal data, then "uses your data everywhere, often time without you knowing or give consent". This is often "hidden in the fine print" as to make it overly complicated for citizens to opt out. One participant claimed that this may be a necessary evil for individuals to accept:

> [AI] overreaches its territory. We share information with AI that we would not share with the personnel in the municipality for example. So, it is a very peculiar relationship where in order to enjoy the security [AI] gives, you have to sacrifice a lot of information from your side as we use it that we would not do otherwise.

Though privacy concerns are felt at the individual level, they also have larger ramifications on society. Participants referred multiple times to the perceived datafication of society, promoting that "data really is the new gold". Thus, to enjoy benefits that AI promises to deliver, society is "unnoticeably sacrificing all aspects of life". The data collected is used by AI, and "lead towards certain answers that could be manipulated as well". This signals a genuine concern cited by every participant – prejudice in AI outcomes. Having access to large amounts of personal data that aid in decisions, participants worried that these decisions increasingly become opaque the more AI is used.

Individuals in society are reduced to sets of categorizable criteria, resulting in a lack of nuance. This nuance is often attributed to the result of human judgment – a skill that AI lacks. Even more worrying, is that this categorisation occurs "behind the scenes, where there are a lot of things going on that people do not know about" and that "you [may be] considered a suspect just based on factual information about yourself". One participant indicated that "it is like shooting with hail". As a result, certain groups of people may be overrepresented in specific datasets, which leads to unfair outcomes. For instance, one participant highlighted that when developing criteria to detect social benefit fraudsters, "certain groups of people are not even in the selection criteria so they will be completely overlooked in principle, which is horrifying".

Moving beyond prejudice, participants discussed three main ways in which the ethical implications of AI can cause breakdowns in the cohesive elements of society. First, as AI applications become ubiquitous in society, this leads to deskilling and de-professionalisation in certain job areas. This may widen the gap between individuals with different levels of education. For instance, one participant recalls how they lost their job due to the implementation of an automation project in their organization. Similarly,

another participant also underscored the need for "awareness that some of the intelligence takes jobs of people that cannot go anywhere else". Second, elements pertaining to digital literacy play an important role, particularly when individuals are users of AI systems. Older individuals are often at risk of being isolated if they do not possess the necessary digital skills. As one participant exclaimed:

> I think that there will be a division between the old, the elderly, people like me. And the new generation. I mean, they probably lived easier with this new development than people our age or a bit older. I feel a sort of resistance myself, but I am a digital fossil.

The aforementioned statement highlights that the gap may not only be the result of a lack of digital literacy, but also due to resistance from older generations to change their established techniques. Third, the access of AI to large amounts of personal data lays the foundation for opportunities relating to personalisation of content, public services, information, etc. Participants consider personalisation as a threat to their "secure democracy" and the cohesion of societal members. If the government is able to personalise services and information based on personal data, then this channels certain societal groups in a one direction while others in a different direction. Thus, citizens are no longer all subjected to the same information, and in the absence of an "overarching view, everybody just can be confirmed in whatever you want by whatever algorithm fits you and your ideas".

In light of these ethical implications of AI use, participants highlighted the current lack of establishment of required safeguards to ensure these concerns are mitigated. Accordingly, AI is attributed this omnipotent quality that is supposedly untouchable. Hence, AI development is seen as a given in society, that cannot be intervened with:

> For me, [AI] it is something mysterious, something that humans do not have control over … is something that we have no impact on, something that works on its own, and independently… It is an unstoppable process, it is emerging, so we have a kind of nothing to do with each other. We can just rely on the technology that is evolving.

This view of AI as something mysterious stirred some concerns amongst participants, who emphasised that not enough oversight exists over the use of AI systems. Whereas citizens do recognise that there is a certain degree of autonomy they have to cede, participants called for safeguards to ensure that this cession is not misused. In this regard, human oversight is advanced as a requirement over AI decisions. Decisions should not all be left to AI systems, but that the *human factor* is important to ensure that nuance in

decisions is considered. Participants underlined that whereas AI systems cannot be held legally responsible if things go wrong, humans can still be put on trial. Thus, the ability to hold a person accountable and culpable for unfair outcomes is an important safeguard required.

Participants did not only call for oversight over the decisions of AI systems, but also their programmers who are deemed isolated from the repercussions of their actions (i.e., role in developing algorithms) on broader society. Participants referred to the "white male programmers programming" who "at the end of the day, have to kind of self-check themselves". To narrow the gap between programmers and society, as well as enhancing the realisation of policy makers on the impacts of their policies, one participant suggested that:

> There should be more connections between programmers and policy makers. That is where the communication falls apart. Maybe that is case. If it us helping policy makers to make more better informed decisions, then I am all for it. But, again, oversight, oversight, oversight [for the programmers]. I do not think that can be overstated enough.

Beyond the role of programmers, participants also attributed certain responsibilities to the government for mitigating the negative ethical implications of AI use in the Dutch public sector. Currently, the current environment of AI use in the public sector is considered as the "Wild West", wherein increasing negative implications are faced by society, and legislative efforts are too late to mitigate most damages. If not addressed, participants fear that their once democratic society will unravel, with more authoritarian practices by a "more closed off government" becoming more apparent. In such a scenario, participants fear that "everyone accepts that they are basically being watched all the time not only by corporations, but also by the government". In addition to worries regarding surveillance, one participant noted worries about decreased capacities of upcoming generations due to AI use. They cited concerns related to "our capabilities going down because of all kinds of innovations…our ability to rationalise or to synthesise [may] go down because AI does it for us".

Thus, in the name of upholding democracy, participants call for control and oversight elements to fall within the realm of the government. In this way, people may be able to influence the actions of democratically elected politicians "who control our data". This is deemed necessary, even if it entails "accepting the bloated public sector with new layers of overseers and ethical committees".

The increasing role of government as an entity that safeguards individuals and society from the negative ethical implications of AI is more necessary now than ever considering the current context. In the absence of the concrete deployment of legislative procedures ahead of emergent negative repercussions of AI use in society, the government is currently not highly regarded amongst citizens. For one, participants had doubts over the true role of the government in advancing their interests as opposed to solely pursuing efficiency gains. Various participants alluded to the fact that:

> Artificial intelligence in the public sector is more and more run like a business [with a] focus on efficiency and delivering value for the best possible efficient standard. So, I see a big similarity in drive in using AI to become more lean, more agile, more cost efficient, instead of really trying to make it something that is helping citizens or that is increasing citizen participation.

Other participants referred to the "corporatizing of the government for focusing on efficiency". In this regard, participants fear that the government is losing sight of its actual goals of advancing citizens' interests. These elements have overall decreased citizens' trust in the government, with reference to prioritising their interests and establishing proper oversight policies and procedures.

Overall, both the positive and negative perceptions of the ethical implications of AI use shape how citizens understand AI in the public sector. Currently, citizens' perceptions are overtly negative, which can be attributed to the lack of legislation and role of government in safeguarding citizens' interests. These negative elements overshadow beneficial uses of AI, which may cause citizens to pre-emptively reject AI use in their public sector.

4.2 Understanding and Prioritisation of Ethical Principles

This theme discusses how Dutch citizens personally understand the meanings behind each of the ethical AI principles, as the principles' definitions remain contested between different documents. The five ethical principles put forward by Floridi et al. (2018) are reported on, alongside the sixth domain of governance incorporated in the Extended AI4People Framework by Ashok et al. (2022). In addition to discussing citizens' understanding of the principles, their prioritisations of the ethical principles are also explored. This provides insights into what Dutch citizens deem important for the use of AI in the public sector, which is heavily influenced by the local context. As such, the principles are elaborated upon in the order of their importance, as suggested by the participants in their discussions.

4.2.1 Autonomy

Participants cited autonomy, explicability and justice as the most crucial principles to prioritise within AI use in the public sector, deeming them equally important. The Asilomar AI Principles (2017) regard autonomy as "humans should choose how and whether to delegate decisions to AI systems, to accomplish human-chosen objectives" (n.p). With reference to this definition, participants heavily indicated that AI decision-making "should be mostly advisory and supporting", rather than giving the final verdict. One participant underscored how augmenting AI decision-making by humans failed within the context of the *toeselagenaffaire*:

> If you then set a standard like [the Dutch government] did, like, okay, we do not have the time or the degree to check everything, so we just put everybody on zero unless otherwise decided, then you are going wrong. No, it should help you to find fraudulent cases earlier. And that still means that humans need to check; everybody is innocent until proven guilty. So, you need to go about it in a very highly integrated way.

This shows how participants value having a "real person" checking AI decisions, to mitigate the emergence of unfair decisions that do not incorporate the "human element" or "nuance in decisions". More than just overseeing AI decisions, participants also highlighted the importance of monitoring procedures throughout the entire process. Such monitoring is crucial for instances where AI is deemed contested, such as in facial recognition applications. As much ambiguity remains regarding how exactly AI operates concretely, the increased monitoring is suggested by participants as an approach to tackle this. If any issues arise throughout the process, adequate "escape routes" must be present, delegating decision-making power back to humans. As such, the need for both monitoring and oversight for AI decisions is highlighted in the following:

> I think that people, the citizens should always have the final say, the final rule over AI. And that is not black and white, I think it is going to be a spectrum of sacrifices. In autonomy, you have to agree what level of citizens' autonomy you are willing to give up, even if you know that some citizens are not going to be happy that designing AI means it is autonomous to make sure decisions over them. And at the highest level, it should always be a way to basically oversee the AI decisions to limit its autonomy.

4.2.2 Explicability

Intertwined with how participants regard autonomy is the principle of explicability. This encompasses two distinct concepts: intelligibility regarding how AI works and accountability which encompasses responsibility for how AI works. Within autonomy, participants underpinned the need for humans to be incorporated in the process, one of the reasons being to ensure that nuance is part of the decision-making process. Participants expressed that an important precondition for incorporating nuance in decisions is to have transparency as to how these decisions were made. Thus, this shows how intelligibility is mostly regarded as relating to transparency. Transparency is deemed important as within the Dutch public sector, information regarding AI systems and how they are used remains ambiguous. As one participant proclaimed:

> … is that the technological possibilities are way ahead of what we know as normal citizens. We do not understand the ethical and the moral implications of what is going on. And we are all like, Wow, this is possible. This is possible, but it is not going in this [direction]. It should develop more. And we are always walking behind. That is what scares me.

The lack of transparency is not only regarding concrete AI applications in the public sector. Participants also expressed that they are unaware of how their data is used and how decisions using AI are made, often describing these things as occurring "behind the scenes, without us knowing". As such, participants emphasise that "transparency is almost vital to the function of AI". Nevertheless, in the current Dutch public sector context, this is not yet the case. In the absence of knowing what goes on behind the scenes, participants place heavy emphasis on accountability for the decisions produced. This relates back to the principle of autonomy, requiring that a human is always in the loop.

The incorporation of humans in the process is vital for accountability, as "if something goes wrong, you could hold somebody responsible for it" and "humans you can put on trial". These aspects show that in cases where miscarriages of justice occur, that would be the result of a lack of human monitoring and oversight, wherein responsibility to a human must be made so that they can be held accountable. In doing so, participants regard the process as still being with the control of the government ("the democracy elect"), which gives them some influence over the process.

Encompassed under accountability is also the idea of proportionality. Various participants proclaimed that in most cases, the government should be held accountable in instances where unfair outcomes emerge. In addition to this, some participants noted that the government is also responsible to respond in a proportional manner if such outcomes are

identified. Taking the example of the *toeslageneaffaire*, the AI system flagged various individuals as potential suspects for fraud. In this regard, it is the government's responsibility to respond in a manner proportional to the severity of the situation. One participant expressed some concerns when stating that "but to collect millions of millions of data of a taxpayer's, innocent taxpayers, to hunt five people down for fraud or something, that is not proportional".

4.2.3 Justice

Participants relate the lack of oversight and transparency in AI-related decision-making with the emergence of unfair outcomes, highlighting the intertwinement of the three ethical AI principles. These unfair outcomes are perceived as inherent to the use of AI since justice is "directly opposite to algorithms. You have to bring down everything to rules, to sets of data and categories, so nuance is not there". In this regard, AI is seen as "limited" since it depends on "facts, and nothing more". Accordingly, the principle of justice within the context of AI is understood by participants as pertaining to discriminatory decision-making, which overrepresents certain groups such as in crime statistics, while ignoring others. As one participant phrased the issue:

> As cases showed, [AI] is very likely to start discriminating against certain groups of people. I will be afraid if you start involving AI into enforcing laws, or try to predict where crime is, I think it is probably going lead to a very discriminatory style of policing and probably people in areas which are not doing quite well will probably be more punished for the same crimes than those people in wealthy areas.

This suggests a recurring view amongst participants that AI use should be minimised when dealing with social matters in the public sector. Another element associated with the principle of justice is the *adaptability* of AI within local contexts. This can be viewed under the header of promoting prosperity, which incorporates attention to local culture and context. Adaptability of AI entails that such systems can incorporate relevant factors in their decision-making, which can bear varying significance in different contexts. As one participant summarised the issue:

> The reason why I think [adaptability] is very important is to think about the masses of people, and we have to consider the factors that are similar and different on a greater scale since AI is being used on a global platform right now. AI is going to hit everybody. We have to pay attention to new aspects. When we talk about the wheelchair for example, how AI will decide who is worthy of or more intensive, let us say treatment or health. Then, you will

> have to figure out a formula of what impacts people's health in a given geographical area – is it the heat, is it hunger? And according to those measurements, you have to apply to a different geographical setting.

Hence, participants regard justice in relation to discriminatory outcomes within the Dutch public sector. Perceptions regarding justice are not only affected by the overrepresentation of certain groups, but underrepresentation of others, leading to biased results and breaking down of social cohesion. Nevertheless, the aspect of adaptability highlights possible room for improvement through considering the needs of citizens in differing contexts. This might address an issue commonly cited by participants regarding ambiguity surrounding terms such as fairness, as "fairness is what is the most important thing to us, is something different than fair to somebody else in a different place".

4.2.4 Non-maleficence

The prioritisation of the non-maleficence principle follows that of autonomy, explicability and justice. This principle incorporates the assumption that AI should *do no harm*, wherein similar sentiments were reported by the participants. In discussions, participants conveyed the emergence of AI-related harm in regard to three main elements: privacy, personalisation and education.

In relation to the first element, participants voiced genuine concern over privacy issues relating to the use of AI in the public sector. Participants noted that inherent to the use of AI, is this aspect of privacy, as can be shown in the following excerpt highlighting a discussion between the participants:

> [Participant 1]: I am wondering the use of AI, has it always to do with privacy? Is it always connected to privacy, to submit to privacy? Or is this always connected using AI? Is it always about privacy, gathering data for citizens?
>
> [Participant 2]: I think it is, yeah.
>
> [Participant 3]: Yeah. In the public sector, of course.
>
> [Participant 2]: The data is really the new gold.

The main issues pertaining to privacy is that data is collected about citizens who lack enough information about how their data is used, and often are not given the opportunity to consent for their data being used. Participants indicated that this poses an ethical issue for the use of AI in the public sector, particularly as the government is portrayed as the

instigator of these processes, ultimately harming citizens rather than advancing their best interests. Participants also expressed concern over their data "coming in the wrong hands, as it can be used against you if there is not enough transparency". To tackle these concerns, participants expressed that AI may uphold the principle of non-maleficence if citizens are able to retain control over their data. In regards to the case discussed (see Appendix, example 02), participants noted how the non-maleficence can be upheld when using AI:

> They do it pretty well because you can submit your own data. Yeah, which is nice because then you can choose for yourself what type of data you submit. And there is nothing straining the relation between the municipality and residents.

As for the second element, participants proclaimed that personalisation relating to AI use results in harm within society, particularly for already vulnerable individuals (e.g., children). With such personalisation, people "do not communicate anymore with each other" as they are channelled in directions that are relevant to them, thus negatively impacting social cohesion. To be able to personalise, citizens are reduced to a set of a quantifiable criteria, which undermines "the human factor" of considering nuance. This allows for citizens to be steered in particular directions while having their own views reinforced. For this, participants called for "opening up" by ensuring that "the information coming in [should be] unfiltered".

The last element pertaining to how AI use can cause harm is related to education. For instance, within the public sector, AI can be used to personalise public services according to citizens' information and stage in life. In doing so, citizens no longer have to excessively search for their choices and sift through alternatives. Although this highlights a simplified example, a few participants expressed concerns over how the use of AI might reduce our abilities to rationalise and synthesise information on our own, as AI does these tasks on our behalf. This results in "programs that become even better than us". Thus, this issue highlights an important aspect, a question posed by another participant: "that is something to consider; to use AI pretty much for everything and how much power it will use. Do you really need AI for everything?".

4.2.5 Beneficence

Beneficence, though highlighted by the participants, was prioritised below the abovementioned principles. Corresponding to AI should "only do good", participants indicated several areas where AI promotes the benefits of citizens and society. Such areas include the using of AI in mobility and infrastructure, to connect previously siloed

organizations and make prediction models based on data provided, particularly relating to environmental issues. In these cases, AI is perceived as speeding up "slow independent manual processes", which otherwise would have taken too long. One participant conveyed the benefits AI use can have in addressing societal issues:

> Because you have hackathons and you get all these really crazy, uh, whiz kids. And that is interesting. Every sector can have a problem; medical, biological, whatever. Lately, there was a woman who was working with illiterate people, and she sort of posed a problem. And then you have these tech nerds and they do not see, for them it is data. They just work on the data. And for them it does not really matter if they are working on something medical or whatever, but they can find a solution.

Despite this view, various other participants strongly urged that AI use should be delimited to non-social elements as "when it comes to certain social things, an AI just simply cannot comprehend the situation properly in my opinion". Accordingly, in the words of another participant, AI should "stay out the social sphere, and limit yourself to environmental problems".

In addition to noting the benefits that AI use can offer, participants also mentioned elements pertaining to its impact on human dignity. The term entails not just ensuring that AI decisions are non-discriminatory as to uphold the dignity of citizens, but also to elements such as job displacement or shifts in society as a consequence of technological use. The digital gap between generations was the most cited concern impacting human dignity, which participants suggested will only become more prevalent as the ubiquity of AI use is more prevalent. When asked if we should only rely on chatbots for customer service, one participant argued that only "in maybe 20 or 25 years, when there is only generations that have been growing up with learn to use it" can AI be considered not exacerbating gaps between citizens. Thus, in the name of upholding dignity, multiple participants noted the need for a non-AI option, particularly when a human-machine interaction is central to the service.

An interesting perspective put forward by one of the participants was in relation to the principle of beneficence incorporating more than just human dignity, but ultimately AI dignity too. Although this statement is not applicable in the current context of AI use in the public sector, this participant spoke of a future where AI becomes "truly aware". In this context, the participant called for respecting AI, for it to be seen as "a friend and a companion, and not just a tool". Although other participants resonated with this statement, they noted that other relevant moral consequences of AI currently take precedence over technical elements.

4.2.6 Governance

Alongside highlighting elements pertaining to the five ethical AI principles as put forward by Floridi et al. (2018), participants also discussed aspects relating to the governance of AI, signifying the relevance of this additional domain as discussed by Ashok et al. (2022). Similar to the definition put forward regarding this domain, participants elaborated upon the regulatory, financial and economic, as well as the individual and societal impacts of AI in their discussions.

Regarding the regulatory context of AI use, participants suggest that in its current context, inadequate regulatory mechanisms are in place, presenting a cause for concern. One participant noted that accidents pertaining to AI use inevitably occur, but that they may be beneficial as these provide the necessary push to adopt legislation regarding the matter and "get it into their cage again, under control". There is the perception that the private sector is bound by much greater rules than the public sector, who participants believe only uphold minimum rules pertaining to the General Data Protection Regulation of the EU. One participant emphasised that due to the lack of transparency over relevant regulation or what is exactly lawful, citizens are forced to take a "leap of faith that the [government] will use [AI] in a positive way". Nevertheless, even if regulation is sought, participants expressed concern over what values will be leading, and how to conceptualise terms such as fairness in an uncontested manner. Political debates are to be expected, but a first step is still required in the matter:

> I think it is going to come down to a clear and defined document that we all have to follow. I think, based on past experience, that seems to be the best way to implement something. If we cannot agree, let us just figure it out and put it in a document and we can disagree later on.

The financial and economic impact of AI use was more negatively perceived, due to the aforementioned concerns of the "corporatizing of the government by focusing on efficiency", causing them to "lose sight of the actual goals". With these views, participants conveyed a decreasing sense of trust in the government, which is no longer acting in the citizens' best interests. Other concerns include citizens' information being "monetized" and "used as a commodity". Thus, rather than upholding public values and safeguarding citizens' interests, the use of AI in the Dutch public sector is associated with the government being run more like a business, which comes at the detriment of their reputation and citizens' perceptions.

Various issues cited under the individual and societal impact of AI use intertwined with the previous ethical AI principles, most notably beneficence. AI use is perceived as

causing fundamental shifts in society, which participants expressed concern and fear over. These shifts relate to the de-professionalisation of certain jobs in society, including AI taking over jobs "of people that cannot go anywhere else". Alongside job displacement, the digital gap between generations is also spotlighted as the older participants emphasised they feel personal resistance to the technology. This resistance stems from the unfamiliarity with AI systems (exacerbated by the lack of transparency), but also perceptions that AI cannot adequately incorporate nuance in its systems and decisions. Considering that older generations did not grow up with similar levels of technology, they are less willing to adapt.

As such, this theme portrayed how Dutch citizens understood and prioritised the ethical AI principles, based on their general knowledge and experiences with how AI is used in the Dutch public sector. Their understandings of the principles are similar to that in the literature, with a few additional interesting findings, such as AI dignity. How participants prioritised the ethical principles sheds light on the local context, highlighting the need for transparency, explicability and justice foremost. This adheres to rationales of first focusing on moral aspects of AI, then to the technical aspects (e.g., regulation). These elements thus provide a glimpse as to what is deemed important in the Dutch context.

4.3 Ethical Tensions

This theme discusses the values prioritised by participants when AI is deployed in the public sector. These prioritisations are influenced by citizens' needs, relative to the local context. As such, the results of this theme provide insight into how high-level principles should be implemented in practice, and priorities for when ethical tensions emerge. Four main ethical tensions were presented to the participants, as put forward by Whittlestone et al. (2019). These tensions were introduced by four fictious scenarios (see Appendix, cases 01-04), wherein participants discussed which values they found leading.

4.3.1 Efficiency vs. Privacy

The first tension refers to issues arising when efficiency benefits are pursued, which are at odds with privacy as large amounts of (personal) data are needed to enhance efficiency. Most participants proclaimed that within the public sector, privacy is more important to uphold, even at the expense of AI efficiency. When asked why they prioritise privacy, one participant highlighted that "I am a bit allergic to the word efficiency". Two main reasons were shared in support of privacy. First, participants expressed genuine concern over the security of their personal data after it is collected, particularly when the data is gathered for social services. A lack of trust in the government to ensure the security of data was advanced by one participated when stating that "I do not trust my government

with the data they will take from me as it would be on the street in ten years". In social services, these fears are heightened as such services are considered "very personal", with privacy perceived as relating "to the dignity of human beings".

Second, the increasing efficiency of AI is seen as coming at the expense of ensuring nuance in decision-making, referred to by participants as the "human element". Individuals seeking public services "all have different stories", with their details unable to be captured adequately by AI. Particularly in relation to social services (e.g., unemployment benefits; see Appendix, case 01), participants exclaimed that the element of nuance is crucial:

> The reason why I would go with upholding privacy is because I would say that it would be very difficult for people with criminal backgrounds or drug use for example, to enter the society, it will be very difficult to find a job and to enter the job market. So, I think that privacy would be the priority in this sense over efficiency because it might eliminate a large group of people from everyday economics and social interactions.

Hence, for instances where AI decision-making is used, participants highlighted that "AI is going to have to make a decision on limited information, it is going to have to be left efficiency because you need to uphold privacy".

Despite this, participants expressed that the efficiency of AI is one of the main reasons behind its use in the public sector, and that it should be pursued. For instance, the use of AI to speed up processes pertaining to the provision of services is deemed as a positive aspect of AI. In this regard, participants underscored the need for the establishment of several safeguards to harvest efficiency gains in a more controlled manner. Several relevant pre-conditions were advanced by the participants, which they currently do not view as established in the Dutch public sector.

One pre-condition noted by the participants is that the data used by AI should be large enough, as that is perceived as enhancing nuance in decision-making. Nevertheless, there should be a boundary for the data that the government is able to collect about certain individuals, which ought to be clearly defined. On its own, "efficiency sometimes can be very careless". Accordingly, citizens should not be asked to provide data that is irrelevant for the process for which AI is used, such as obtaining information on one's ethnicity when applying for benefits. Another pre-condition is creating a committee or a controlling body which monitors the outcomes of AI, rather than blindly following AI-made decisions. This relates to the importance of oversight. Upholding these conditions is seen

as a positive step by participants to sharing their data more willingly, which serves to boost the efficiency of AI:

> But if we just can work out the issue of how data is being used, consent forms, whatever it is, again we need to have tighter regulations on these procedures. If we can control that part, I think the answer becomes more geared towards efficiency.

This highlights the current lack of safeguards for citizens' personal data, leading to concerns regarding the security of this data and hence, the prioritisation of privacy.

4.3.2 Accurate Predictions vs. Fair Treatment

The second tension refers to the conflict arising when using an accurate AI system, at the expense of having representative data for the population that enhance fair treatment in outcomes. Participants were split on which value is more important to pursue within this tension. On the one hand, some participants perceived this tension as relating to issues of equality and representation. In this regard, this group wanted to ensure that AI systems, though supposedly accurate, should only be used when it can be "made sure that both groups are treated in the same way". When elaborating on whether accuracy alone in AI systems should be pursued, one participant claimed:

> Absolutely not. I do not think accuracy is a factor here, just the purpose of such system is already inherently discriminatory. You are already sidestepping an obstacle and you are using the system in a discriminatory way. Again, I am not saying that human judgment is perfect, you could have discriminatory practices in parole hearings for example. I do not think that can be avoided, but if you are putting the two things on the table, I would go for the human judgement on that one.

In this regard, participants emphasised that in the absence of representative data for all groups, decisions made by AI should not be considered, and human judgment solely relied upon. The latter's judgement, though not flawless, is perceived less discriminatory than using AI systems, and has the added factor of considering nuance in decisions.

On the other hand, participants also underscored that the accuracy of the AI system cannot be ignored, sometimes even at the expense of the system not resulting in fair treatment. If given the choice between using the accurate AI system but that representativeness cannot be ensured, participants voiced support towards using the system, even if it solely works for specific groups of people. The rationale behind this argument is that an accurate AI system can provide sped up processes and outcomes for a subset of people, thus the

system should not be completely eliminated. In this regard, enough nuance is expected to be in the data for the subset of people it serves, promoting its use. For the other groups of people where representative data is not available, participants suggested to just "figure it out" and "find something else", including the introduction of "customised" and "personalised decision-making". This is highlighted in the following:

> Assuming there is still a final decision of a [human], I think you can keep using the system even though it excludes a group of people because at the end of the day that helps to get more people, it improves the well-being of a number of people. You can even call it affective altruism. It is not entirely fair to the group of people who are not getting to speed up the process. So, for me in this case it does not harm the consideration that if there is already right now a way to speed up the process for a large group of people then start with that and see if you can apply that more universally.

As such, emphasis is still placed on ensuring that even for groups where AI systems are used, they should be supplemented with human oversight. Accordingly, the aforementioned perceptions hold in instances where AI is part of the decision-making process, and not the sole decision-making entity. As for the other groups where representative data is not present, participants called for prioritising human decision-making when AI decision-making is inadequate. Moreover, one participant expressed that in such cases, this pitfall gives an incentive to gather additional data and fill in gaps, which serves both short and long-term benefits.

4.3.3 Personalisation vs. Solidarity

The third tension refers to the predicament between ensuring that every citizen is treated equally and applying filtering tactics for citizens relative to the information and services they receive. The former relates to the element of solidarity, while the latter incorporates personalisation. Similar to the previous tension, participants were also split in reference to which value they deemed more important, presenting relevant arguments for both sides.

In regard to solidarity, some participants noted that it is crucial to uphold, particularly in a democratic, pluralistic society. In this sense, it is not the role of the government nor PSOs to decide on what information or services citizens need to be notified about. Accordingly, citizens reserve the right to be made aware of what is going on within society, whether they are interested or involved in a specific topic. For instance, the scenario for which this tension referred to introduced a potential municipal meeting regarding sustainability, and inquired who invitations should be sent to (see Appendix, case 03). Those in favour of solidarity responded that these invitations can "just be put in

everybody's door", otherwise "you are already profiling". This allows ample opportunity for discussions between social circles (e.g., neighbors). In doing so, this provides all citizens the opportunity to participate, and potentially even gain knowledge regarding a topic they have not previously engaged in as "you do not know what you miss".

Nevertheless, concerns were voiced over the lack of efficiency associated with solidarity in this context. Citizens are already overly stimulated with the amount of information they receive, with one participant noting that "I think there is so much coming at citizens in terms of communication and information overload". In addition to this, sending out invitations to everyone "will cost too much money", often taxpayers' money, without guarantee that all those invited will attend. Even if a greater number of citizens attend, they may not be interested enough in the topic to actively participate and share their views, which can stifle discussions around proposed solutions and next steps, and result in longer meeting times. As such, participants note that such personalisation is more conducive to small topic matters, as larger social issues ought to be discussed with all members in society affected:

> I think only people who previously shown interest, at least when it comes to small things, mostly advanced interest in giving their opinion about all things related to governments. So, I think it would be just a waste of time to insist on having everybody know about this essentially.

Moreover, sending invitations to everyone also entails sending to those who have recently moved into new areas/communities for which they were previously unfamiliar with. This raises concerns over newcomers sharing opinions on matters impacting their community, where they themselves cannot be considered representative of the community's general perceptions. Sometimes, newcomers "need months, even years to grasp the social settings of a closed environment", and thus need "time and lots of investment to understand where they actually are". Although opposed to solidarity, one participant indicated that this approach is "not discriminatory, I do not think it is a favouritism-type scenario, it is personalised".

As such, participants recognise the value of ensuring that all relevant information/services towards a specific community/society need to be delivered to its citizens as their right. However, this may be too costly or inefficient for the result achieved. In some instances, such as with newcomers and people who have not expressed interest, solidarity is not sought and the personalised effects of AI use can prove beneficial.

4.3.4 Automation vs. Dignity

The last tension is considered core to debates surrounding AI – to what extent should AI be used and what should be left within the realm of human control, and by extension the dignity of humans. In the context of the public sector, the use of AI promises efficiency benefits, but often at the expense of processes citizens are used to, and even consider crucial to their perceptions on how society functions. This dilemma is apparent in discussions between the participants. The use of AI, for instance to respond to general inquiries and retrieve information, is beneficial and reduces the need for performing mundane, routine tasks:

> You just want to get over with your problems, you just want to get over with whatever difficulty you are facing. I think that it is a great way. It is also saves so much energy for public workers who do not have to listen to unreasonable arguments…So, you can sort these things out and go through with these in a lot easier ways with AI.

Despite this, the lack of nuance and room for discussions with AI is negatively perceived, and leads to much hesitation over the use of AI in social contexts (e.g., AI-enabled government chatbots). AI is seen as conducive to offering "yes or no answers" or "very limited open answers" instead of providing well-rounded responses considering multiple aspects. Citizens do not want to feel that their inquiries can be categorised and their answers pre-determined, even if that is indeed the case. As one participant noted, "I think it is really in poor taste that there are no real people at certain spots, I really miss them". As such, this issue can be summed by the following statement:

> For the public sector you should always keep humans beside of [AI] for those that want. But as long as AI cannot perfectly emulate humans then those people who are older or are not familiar with technology will struggle with that. So, you need to keep the humans.

This touches upon a second concern voiced by participants – the digital gap. As one participant claimed, the use of AI "is a generation thing I am afraid". This entails that the use of AI is not inclusive to the needs of all groups in society, including older generations and those with lower education. If AI were to replace service employees, one participant highlighted that "I hope that I do not live then anymore, it is really sad". The same participant expressed that the exclusion of people "is my phobia", representing a repeating concern within the discussions.

Nevertheless, participants suggested that these perceptions are due to the current developments and uses of AI, which are ever-evolving. Current and upcoming generations grew up with technology, and are considered more adept to witnessing technological changes in society. One participant noted an instance where a major change was accepted within society:

> For example, if we think of maybe ten years ago, I do not know when we could buy a ticket at a at a railway station. Yeah. And then the oldest personnel got gone and there is only automata. And we were like, oh my God, this is what is happening. And it is oh, impossible. And old people cannot travel anymore. And now it is completely normal and nobody misses this kiosk with a person. But I am just thinking it also very difficult to step away from your reality right now.

Although AI is not currently at the level where it can replace employees and be positively perceived in this regard, participants proclaimed that that point is likely to be reached in the future.

In conclusion, this section elaborated upon the general perceptions of Dutch citizens over the use of AI in the public sector. Although the benefits of AI use were duly noted and the gains from that acknowledged, negative perceptions continue to dominate discussions. The overly negative perceptions influence the prioritisation of ethical principles amongst citizens, citing transparency, explicability and justice as the most relevant to uphold. This highlights a trend wherein moral consequences of AI use should first be considered, followed by their technical elements. Participants also shed light on what values are leading within society, noting important values to pursue when ethical tensions and conflicts inevitably emerge. All the aforementioned themes are regarded as reflecting elements relating to the Dutch local context, which presents peculiar aspects in a unique setting. As such, the results from the participants are analysed relative to the Dutch context, and their consequences on society discussed.

5 Discussion

The Netherlands has recently been pushing an expansive agenda for strengthening the role of ethical safeguards within AI use, both within the public and private sectors. The agenda is considered relevant due to the ubiquitous use of AI in the Netherlands, as evidenced by the country's advanced position in both the DESI and AI Indices. These ethical safeguards are regarded as being established ahead of the implementation of the AI Act by the EU, which is a stringent regulation promoting the responsible use of AI through strict conditions and procedures.

Nevertheless, the Netherlands was previously ill-regarded in relation to the AI use in the public sector. Considered the worst AI scandal in Europe, the *toeslagenaffaire* was heavily covered in the media as a warning story against the incorporation of AI systems into the public sphere. Such media coverage was extensive, and heavily influenced Dutch citizens' perceptions of AI use. This is exacerbated by the low level of public trust in the current government, which can be seen in the all-time low approval rating of 20 percent – the lowest in a decade (Reuters 2023). As such, a complex environment currently exists in the Netherlands – a country which is ahead in innovation, digitalisation and provision of public services, but within the backdrop of declining public trust and wavering citizen acceptance of AI use. This highlights relevant contextual elements that influence participants' answers and perceptions over the ethical implications of AI use.

5.1 General Perceptions

The use of AI in the Dutch public sector has ramifications on broader society. Citizens are impacted by the use of AI within various domains of their lives, including healthcare, education and employment. Despite being impacted by the implications of AI use and its increasing incorporation in society, a lack of understanding of what AI entails persists. This conceptual ambiguity is a heavily cited concern in AI research (e.g., Wirtz et al. 2019), with a common consensus regarding defining AI in research still not established.

Beyond research, there is a current lack of sufficient sources for Dutch citizens to draw from to further their understanding about AI, particularly regarding the public sector. News media coverage regarding concrete applications of AI in the Netherlands are scarce, and mostly refer to the *toeslagenaffaire* as a cautionary tale against AI use (e.g., Heikkilä 2022). Even when AI is discussed in the media, various terms such as machine learning, automation, and AI are often grouped together or used interchangeably that it is hard to discern the meaning behind each term (e.g., Aldane 2023). The lack of media coverage may be the result of the lack of communication by the Dutch government and

municipalities regarding the incorporation of AI in their processes and services, which further contributes to ambiguity regarding the concept.

These patterns are in line with the argument put forward by Maragno et al. (2022) regarding the two-fold nature of the issue of defining AI – the term is used too broadly to be understood and the lack of communication regarding (successful) applications of AI. As a response, organizations like the World Economic Forum advanced the need for universal AI literacy, prompting the Netherlands AI Coalition to lead an AI Parade wherein dialogues with citizens and demystification of AI occur. The goal was the "sharing of a complete story in a qualitative and human-centred way", and aided the Coalition in researching citizens' perspectives and knowledge on AI, alongside contextualising relevant benefits and limitations (Loohuis 2022, n.p.).

The general perceptions of Dutch citizens towards AI use are aligned with the two conflicting dimensions put forward by Ingrams et al. (2022) as impacting citizens' perceptions – the instrumental and value-based dimensions. Under the instrumental dimension, Dutch citizens positively acknowledge the benefits provided by AI in terms of efficiency, cost savings and improving accuracy. In government domains such as environmental protection and the provision of general public services, citizens express support for the use of AI as it offers more accurate predictions with the former and enhances the speed at which public services can be provided with the latter. These perceptions align with research regarding positive benefits of AI use in relation to decreasing costs and enhancing decision-making accuracy (e.g., Miller and Keiser 2021).

Dutch citizens also advance benefits of AI use beyond the often cited efficiency and cost savings in research. AI use is suggested for tackling societal wicked issues. For instance, loneliness is now an urgent issue in the Netherlands, impacting 40 percent of Dutch people aged 15 and older (NOS 2022). In this regard, the public sector can expand on the opportunities of AI to tackle the issue, such as through widespread deployment of a conversational chatbot, particularly amongst the elderly who are most isolated.

Moreover, AI can be utilised for increasing transparency surrounding politics by providing decision support. This aids policy makers with required insights and predictions for long-term oriented decisions and in addressing complex societal challenges. Although this overlaps with accuracy and efficiency claims, there is a broader benefit of enhanced transparency in decision-making, which is particularly relevant in the context of decreasing public trust in the government. In this regard, AI use is associated with delivering a *government of the future* more responsive to citizens' needs and enhancing trust through heightened transparency (Margetts 2022).

The value-based dimension is more salient in citizens' perceptions of AI use in the Dutch public sector. This dimension is heightened considering the critical coverage of the ethical implications of AI use associated with the *toeselagenaffaire*. Such coverage negatively frames the ethical implications of AI use, impacting citizens' opinions and acceptance of the technology (Ouchchy et al. 2020). Despite more recent coverage focusing on the benefits of AI use and the establishment of ethical safeguards in the Netherlands, the negative frame persists and continues to dominate perceptions towards AI use.

Dutch citizens pose similar concerns about AI use as those put forward by Kieslich et al. (2022). The lack of transparency and oversight in the public sector is associated with AI black box decision-making, which Thiebes et al. (2021) argues negatively impact citizens' perceptions. This can be seen in responses highlighting issues regarding privacy, unfair outcomes, the lack of human element and nuance. These issues are important to address, considering the loss of trust in the government and increasing heterogeneity in the country.

Regarding the former, privacy concerns are dominant due to the recurrence of data scandals in the Netherlands, alluding to structural privacy issues (Mooi 2021). As such, citizens hold perceptions that their data is unsecure and can be used against them. This contradicts the view of Willems et al. (2022), who argue that the government should uphold public sector ethos. With the latter, the increasing heterogeneity in the Netherlands entails that the public sector has more diverse citizens to attend to. Unfair outcomes which disproportionately impact certain groups are already recognised as a problem, particularly in the context of the *toeslagenaffaire* (Heikkilä 2022). Alongside increasing feelings of discrimination by citizens with a migration background (Dagevos et al. 2022), AI use is construed as a system that replicates societal biases and is negatively perceived, in line with the findings of Miller and Keiser (2021).

These negative perceptions are attributed to the lack of adoption of an ethical approach towards AI in the Netherlands prior to the *toeslagenaffaire* and its ethical implications on society. This resulted in missed benefits from leveraging AI and costly mistakes. Hence, Dutch citizens pose similar sentiments to Floridi et al. (2018), emphasising the urgency of an ethical approach, particularly within the public sector where citizens are impacted. The increased government involvement in regards to setting out safeguards and regulations is called for to ensure that the benefits of AI are reaped. In addition to this, Dutch citizens support the approach of a shift to *micro ethics* suggested by Hagendroff et al. (2020). More emphasis and safeguards are called for to be placed on AI developers, considering their role in societal impacts. These points are currently being addressed (ECNL 2022), indicating the start of an ethical shift in the use of AI in the Netherlands.

5.2 Ethical AI Principles

Based on the ethical implications of AI use in the Dutch public sector, citizens discerned the need for an ethical approach, and also formed their understanding of the ethical principles required for such an approach. Citizens' perceptions of the ethical implications aid in setting priorities between the ethical AI principles that are greater aligned with the Dutch local context. In addition to this, the prioritisations provide insight into core values that citizens deem most important. The local context is heavily influenced by the aftermath of the *toeslagenaffaire*, providing empirical evidence for the need of an ethical approach and serves to discern citizens' perceptions on priority areas and values within the public sector.

Autonomy, explicability and justice are considered the most crucial ethical principles to safeguard and prioritise in AI use. This relates to elements regarding the *toeslagenaffaire*, wherein AI decisions were perceived as lacking human oversight, transparency and accountability, leading to unfair outcomes accordingly (Amnesty 2021). As such, this highlights that Dutch citizens hold similar understandings of the ethical AI principles as those put forward in research, particularly Floridi et al. (2018). Autonomy is associated with the overdependence on AI decision-making without sufficient oversight and incorporation of nuance, explicability with the inadequate transparency and accountability in the public sector, and justice with the emergence of discriminatory outcomes. Similar to the argument of Floridi et al. (2018), Dutch citizens also note the intertwinement between the ethical AI principles, realising that pursuing one principle has implications on the others. Accordingly, when referring to the *toeselagenaffaire*, citizens underscore the need for stifling the autonomy of AI considering the lack of transparency surrounding how it works, entailing that human oversight incorporates more nuance and likely to enhance the emergence of fair outcomes.

Dutch citizens' understandings of these ethical principles also incorporate novel aspects. For instance, the concept of proportionality is deemed relevant for accountability under the explicability principle, despite this aspect not discussed in AI research. Karliuk (2022) argues that although proportionality is recognised in law, it should be considered in AI ethics as it "allows to identify whether a legal act of the organization (1) pursues a legitimate aim (suitability or appropriateness), (2) does not go beyond what is necessary to achieve such an aim (necessity) and (3) does not impose excessive burden upon the individual" (p. 3). This relationship between means and ends is relevant within the AI context, as shown by the disproportionate actions undertaken by the Dutch government in the *toesleagenaffaire* against presumed fraudsters. Furthermore, in agreement with the critiques of ethical AI principles by Mittelstadt (2019), the adaptability of AI to local

contexts is considered necessary when making normative decisions, and is suggested to be incorporated under the principle of justice. In light of the increasing adoption of AI worldwide, Dutch citizens want to ensure that local elements are considered, impacting the outcomes of AI decisions.

Non-maleficence, the principle incorporating the assumption that AI should *do no harm*, is perceived by citizens as mainly pertaining to privacy. Fears surrounding the privacy of their data are held by Dutch citizens, heightened by scandals including the *toeselagenaffaire*, the selling of unlawful data by the Chamber of Commerce for €3.3m (Boztas 2021), alongside the trade of personal medical data during the COVID-19 pandemic (Mooi 2021). These views are compounded by strained public trust in the government. Nevertheless, non-maleficence is ranked below the principles of autonomy, explicability and justice. A similar prioritisation was reported by Kieslich et al. (2022) when evaluating German citizens' perceptions of ethical AI principles.

Beyond privacy, Dutch citizens express concerns regarding AI omnipresence, personalisation and education. Corresponding to what Floridi et al. (2018) posited, the ever-evolving self-development of AI is seen as inevitable, wherein its pace of development cannot be controlled. This is considered at odds with the assumption of *do no harm* as citizens conceive that AI encroaches on their daily lives, yet lack the ability to consent to its use. The ethics acceptability of AI is thus impacted, as the first criteria of informed consent by Taebi (2017) was not upheld. Moreover, although personalisation is widely regarded as a benefit of AI use, Dutch citizens incorporate it under non-maleficence. Following this principle, AI is to be understood as limiting harm through allowing autonomy of citizens and not nudging in specific directions.

In relation to education, Dutch citizens deem it important to ensure that AI does not adversely impact learning and education. With rising reports of AI applications (e.g., ChatGPT) disrupting education (e.g., Heaven 2023), opportunities on how to capitalise on these technologies for good are promoted. These concerns are also echoed by Zhai et al. (2021) in their review of AI in education, highlighting the potential negative impact of AI in learning due to students' ability to delegate process-intensive work.

Beneficence is aligned with the instrumental-based dimension advanced by Ingrams et al. (2022). Dutch citizens acknowledge AI as advancing the common good in matters pertaining to predictions and accurate decision-making. Hence, the principle of beneficence is associated with the AI use to contribute to efficiency and knowledge, albeit outside the social realm wherein perceptions of the negative ethical implications of AI use are heightened. This is in line with findings from Gesk and Leyer (2022), who argue

that citizens express more support for AI use in the provision of general public services, as they do not go beyond the individual level nor facilitate citizen-AI interactions.

Citizens consider beneficence as incorporating sustainability. Within the ten functions of government, environmental protection is most associated with beneficence, with beliefs that AI can provide opportunities in this domain. Upholding human dignity under beneficence is also discussed as a priority. Considering that *300 million jobs could be affected by the latest wave of AI* worldwide (Toh 2023), Dutch citizens express concerns regarding job displacement and shifts in society due to technological use. Although this may be an inevitable outcome of AI use, citizens have relevant contextual knowledge and want their opinions considered, which Taebi (2017) argues as a relevant criteria for assessing ethics acceptability. The incorporation of citizens' knowledge can aid in identifying opportunities where AI use is beneficial and providing insight regarding how societal shifts are experienced. AI dignity is also highlighted in relation to beneficence, anticipating the rise of a *truly aware* AI. This is associated with the Kantian idea that dignity is derived from autonomy (White 2011), which is relevant as AI is becoming increasingly autonomous. Nevertheless, when reflecting on the topic of whether robots can have dignity, Krämer (2020) argues that AI cannot be regarded as possessing dignity.

The incorporation of the additional governance domain advanced by Ashok et al. (2022) is relevant as Dutch citizens incorporate elements pertaining to the regulatory, financial and economic, as well as the individual and societal impacts of AI use. The establishment of regulatory mechanisms on AI use is considered a priority under governance, despite concerns about regulation keeping up with AI growth (e.g., Gerrish and Morrison 2020). The financial and economic impact of AI is associated with the so-called corporatisation of the government. As evidenced by the scandal relating to unlawful data selling by a Dutch PSO, citizens express a decreasing sense of trust and public values. De-professionalisation in certain jobs and the digital gap are the two main societal impacts underscored for AI use. With the latter, although the Netherlands is top ranked in the EU for digital skills (CBS 2020), a lack of willingness to adapt to AI is expressed by the older generations, which is also supported in research (e.g., Knowles and Hanson 2018). Concerns regarding the potential of a heightened digital gap between generations are expressed, conveying the need to identify and address significant societal impacts of AI.

Dutch citizens' understanding and subsequent prioritisation of the ethical AI principles are closely aligned with the definitions put forward in research (e.g., Floridi et al. 2018) and highlight similar prioritisations to citizens in other Western countries (e.g., Kieslich et al. 2022). Hence, the first critique posed against ethical AI principles in relation to their vagueness and ambiguity not resulting in a common understanding may not be relevant

in the Dutch context. Whereas Dutch citizens express some novel aspects in relation to the principles, their understanding of the principles incorporate key elements of the definitions advanced in research. Nevertheless, the second critique relating to the technical implementation of normative concepts persists. Dutch citizens hold concerns over how normative concepts can be set into technical rules while upholding the various moral and cultural values in society. This underpins the need for upholding the second criteria for assessing ethics acceptability, which pertains to the need for pluralism (Taebi 2017). Incorporating citizens' perceptions as people impacted by AI systems promotes the pursuit of the human-centric goal, aligning with the Netherland's AI goals.

5.3 Ethical Tensions

For the Dutch public sector to adopt a tension-focused approach in the implementation of the ethical AI principles, consideration for elements pertaining to the social and cultural contexts is emphasised by research (e.g., Greene et al. 2019) and citizens. By highlighting which values are leading in distinct scenarios, citizens can provide relevant insights that enhance the human-centric goal of AI and are applicable to the local context. This can serve as a means to address the second critique against the ethical AI principles of the principles-implementation gap (Munn 2022).

Regarding the first ethical tension advanced by Whittlestone et al. (2019), privacy is deemed more crucial to uphold for AI use in the Dutch public sector. Concerns regarding the theft of citizens' data are prominent considering the various privacy scandals in the public sector, compounded with the low level of trust in the government. These concerns are heightened when AI is used for social services, considering the discriminatory outcomes highlighted within the *toeselagenaffaire* and belief that privacy is related to human dignity. This is in line with citizens' fundamental right of privacy, which underpins dignity and values such as freedom of speech (Floridi 2016). Although efficiency gains relating to AI use promote their adoption, Dutch citizens do not currently consider that adequate privacy safeguards are established. This highlights that action is needed to improve this aspect before efficiency can be prioritised.

The second tension between the values of fairness and accuracy in AI use indicates diverging opinions between Dutch citizens. Whereas with the former citizens emphasise the need for representative data and equal treatment in society, efficiency gains relative to time and cost savings are highly acknowledged by citizens promoting the latter value. In regard to the ethical AI principles, this can be considered a tension between justice and beneficence (Floridi et al. 2018). Accordingly, such diverging opinions contradict citizens' higher ranking of the justice principle over beneficence.

This provides insight that priorities between values can differ according to specific AI use cases, which is also emphasised by Whittlestone et al. (2019). As Dutch citizens' perceptions of the ethical implications of AI use are impacted by the *toeselagenaffaire*, the principle of justice is recognised as more important to uphold in society. Justice is prioritised as the AI system in this case was nationally deployed, thus the absence of representative data and fair outcomes can have large, negative societal impacts. Wherein under the use case presented for this tension (see Appendix, case 02), citizens promoting accuracy did so under the assumption that a smaller subset of individuals are impacted by the AI use. Thus, within a small-scale setting where human oversight is feasible, AI can promote accuracy while sufficiently monitored. This argument is also present in research, highlighting that AI benefits are hard to achieve at scale, considering elements such as larger risk management (Cam et al. 2019).

Divergent opinions regarding which value to prioritise is also apparent in the third tension between personalisation and solidarity. In support of solidary, Dutch citizens underscore the importance of equal access to information impacting them, as should be the standard in a democratic, pluralistic society. Having equal knowledge on relevant matters can encourage debates and promote public participation, underpinning the importance of solidarity in public sector communication (Luomo-aho and Canel 2020). Calls for incorporating solidarity as an ethical AI principle were advanced by Luengo-Oroz (2019) to promote "sharing the prosperity created by AI, implementing mechanisms to redistribute the augmentation of productivity for all, and sharing the burdens" (p. 1). Nevertheless, Dutch citizens also point out the benefits of personalisation in reducing cognitive efforts for information retrieval. Lee et al. (2020) show how citizens are overly stimulated due to the intake of large volumes of information, impacting their interactions with the government. In this regard, personalisation can serve to reduce both cognitive loads and costs associated with information sharing.

To balance both values in the Netherlands, Dutch citizens promote the view that personalisation is conducive to small topic matters (e.g., general discussions), wherein absences would not gravely impact citizens and relevant information can otherwise be retrieved. Conversely, Dutch citizens do not support personalisation for the provision of or the enforcement of public services, due to the reliance on large datasets of personal data, which is also discussed in research regarding the Netherlands (e.g., van Veenstra et al. 2021). Such data previously discriminated against vulnerable groups, highlighting the re-prioritisation of solidarity in such instances. Although the personalisation of services is considered still under development, this can provide insight into how to develop the use of data analytics in a manner consistent with upholding public values.

Promoting values pertaining to human dignity is prioritised by Dutch citizens regarding the last tension between automation and dignity. Similar to the arguments advanced by Whittlestone et al. (2019), Dutch citizens hold concerns related to widespread de-skilling, weakening of existing practices and exacerbation of social cohesion decline if automation is solely pursued by the government. Relevant to the scenario discussed under this tension (see Appendix, case 04) are three main criteria of assessing the effectiveness of chatbot performance, as suggested by Chaves and Gerosa (2018) – conversational intelligence, social intelligence and personification. Whereas Dutch citizens perceive the conversational intelligence of AI as adequate considering its ability to recognise and categorise speech, the latter elements hinder citizens' acceptance of the adoption of AI in social spheres. AI development is not considered at the level where social intelligence is met through the recognition of citizens' psychological and emotional states, nor is personification as chatbots with their perceived limited answers do not replicate human complexity and nuance.

These views can be interpreted using the findings of Aoki (2020) regarding Japanese citizens' trust in public AI chatbots, which varied based on its application area. In specific areas, citizens hold high expectations for the information they receive, wherein responses "must provide enquirers with the information they want, employ situational judgement, and communicate with them in a socially proper and empathetic manner" (p. 9). Support for AI chatbots in these instances are lower than in other application areas, where AI can aid with information retrieval whilst not requiring high social skills. This is consistent with the previously expressed views of Dutch citizens wherein the value pursued is dependent on specific scenarios.

As such, reflecting on Dutch citizens' prioritisations within ethical tensions provides insight that can enhance the implementation and acceptance of AI in society. Particular consideration is given to the intertwinement between ethical principles, the need to balance different interests in society, alongside considering both the short and long-term impacts of AI use. Whittlestone et al. (2019) highlight how public engagement is necessary for providing insights on how to tackle tensions whilst representing stakeholders' diverse interests with rigour. Incorporating Dutch citizens' considerations when addressing tensions serves the dual purpose of identifying leading values for AI applications, and providing democratic legitimacy for PSOs with new AI implementations.

6 Conclusion

Resembling various countries worldwide, the Netherlands is promoting AI use in the public sector in pursuit of efficiency benefits and tackling rising societal expectations. Nevertheless, both research and citizens advance support for the view that AI use should not only target output optimisation, but also incorporate crucial safeguards on AI that protect important societal values. The adherence to an ethical approach for AI use is thus called for in the Netherlands, considering previous scandals and the emergence of costly mistakes negatively impacting society and the public's acceptance of AI use.

Although critiques against ethical frameworks and principles are established in literature, they provide relevant guidelines for the implementation of new technologies, without which the emergence of negative societal consequences are unavoidable. Crucially, these frameworks underpin the importance of incorporating citizens' perspectives within the development and design of AI, despite the identification of these perspectives largely missing practically (i.e., in the Netherlands) and academically. As such, this research provides exploratory insights on Dutch citizens' perspectives on the ethical implications of AI use, emphasising pertinent considerations for AI use in the local context.

Regarding the main research question, Dutch citizens' perceptions of the ethical implications of AI use in the public sector are complex, shaped by a context of advanced AI development, declining public trust and experiences of unethical use. In general, Dutch citizens appreciate the instrumental benefits of AI in public services such as improving accuracy, while also acknowledging its potential to tackle complex societal issues such as loneliness. Despite this, value-based concerns dominate citizens' perceptions. High-profile cases such as the *toeselagenaffaire* negatively frame the ethical implications of AI, overshadowing its potential benefits. Dutch citizens worry about privacy issues, unfair outcomes, and the loss of the human element, alongside concerns about AI as a potential tool for discrimination. These worries are exacerbated by the lack of understanding and transparency regarding AI and its use cases.

For the first sub-question, Dutch citizens show a comprehensive understanding of the ethical AI principles, aligning closely with academic definitions and prioritisations seen in other Western countries. Citizens also express unique perspectives, such as the concepts of proportionality and AI dignity under the justice and beneficence principles respectively. Furthermore, aspects relevant to the additional governance domain are advanced, with citizens considering this domain intertwined with their understanding of the ethical AI principles. The principles of autonomy, explicability and justice are prioritised, though emphasis is still placed by citizens on the remaining ethical principles. This prioritisation reflects growing concerns about the lack of oversight, transparency and

accountability in AI use, alongside the emergence of discriminatory outcomes, all of which were evident in the *toeslagenaffaire*. Hence, Dutch citizens call for strengthening the practical implementation of normative values through a pluralistic approach, consistent with critiques against the ethical AI principles.

For the second sub-question, Dutch citizens' value priorities within ethical tensions offer invaluable insights for enhancing the acceptance and integration of AI in society. Whereas there is a consensus between citizens on the promotion of privacy and human dignity, diverging opinions are held about the tensions between fairness-accuracy and solidarity-personalisation. The divergence in the former tension contradicts citizens' prioritisation of the justice principle. Nevertheless, this indicates that value priorities can shift depending on specific AI use scenarios. These insights underscore the necessity to consider the balance between values, pluralistic societal interests, and the short and long-term impacts of AI use in each application. Further, they accentuate the importance of public engagement in the discourse on AI applications, ensuring democratic legitimacy and identifying guiding values.

As such, the perceptions of Dutch citizens indicate that instrumental-based benefits should not be solely pursued until concerns under the value-based dimension are addressed. Given the legacy of past ethical failings, Dutch citizens underscore the urgency of adopting an ethical approach to AI. Autonomy, explicability and justice are deemed crucial principles to prioritise in the public sector, alongside calls for more robust safeguards and regulations from the government. The Netherlands has begun responding to these demands, suggesting the beginning of greater ethically-aware AI practices in the country.

6.1 Theoretical Implications

This research contributes to academic literature relating to AI and ethics in various ways. First, the operationalisation of the ethical AI principles is complementary to current understandings of these principles. Hence, contrary to the first main critique of the ethical AI principles regarding their vagueness and ambiguity, Dutch citizens show an expansive understanding of the principles. Second, novel aspects regarding the ethical AI principles are expressed by Dutch citizens. Such aspects can expand on current definitions of the ethical AI principles, or even promote the addition of new principles for a more expansive ethical scope.

Third, the relevance of the governance domain is shown, suggesting that the use of the Extended AI4People Framework better encapsulates citizens' perceptions of the ethical implications of AI use in the public sector. The use of this domain as an additional ethical

AI principle is promoted, highlighting its intertwinement with AI ethics understandings. Fourth, evidence supporting the second critique of the principles-implementation gap is shown, indicating the relevance of and the need to address this critique to bolster the social acceptance of AI. Fifth, prioritisations of the ethical AI principles are not static, with different values leading in different applications. This furthers the understanding of principles as representing values that hold dynamic degrees of support and are impacted by the specific context.

Hence, this research addresses a large and pertinent gap in current AI research – the lack of inclusion of citizens' perspectives. More than ever, citizens want to feel that their voices are heard in matters that have direct implications on both them and the wider society. This is regarded a necessary step to enhance citizens' social acceptance of and trust in AI.

6.2 Practical Implications

The practical implications of this research are manifold. First, presenting citizens' understandings of the ethical AI principles and their prioritisations can aid in the formulation of policies regarding AI that are greater aligned with societal values and expectations. This research can guide Dutch policy makers on where to focus their efforts, especially regarding the prioritisation of principles and the identification of leading values when ethical tensions emerge. By addressing citizens' concerns proactively, the public sector can work to increase public trust and acceptance of AI. Furthermore, transparency about how these issues are addressed can foster trust in AI systems and their use in the public sector, underpinning the need for greater communication and more expansive media coverage on the matter to actively engage with citizens. This enhances pursuits towards the human-centric goal of AI.

Second, insights from this research indicate the need for greater scrutiny on the developers and implementers of AI systems. The outcomes of AI have a widespread impact, yet developers are regarded as isolated from the repercussions of their actions on broader society. Accordingly, safeguards to ensure that AI developers and implementers are acting in the public's best interested are strongly recommended. This incorporates integrating diverse stakeholders' concerns into the design process, underscoring the importance of involving the public in discussions around AI and its ethical implications. Participatory approaches in policy-making are thus promoted, ensuring a diverse set of voices are heard and bolstering democratic legitimacy.

Third, this research contributes to the broader discourse on AI ethics, emphasising the need to align the usage and outputs of AI with societal values and ethical principles. The

values and ethical principles deemed leading by citizens are dynamic, and often dependent on the specific contexts of AI use. Ethical considerations pertaining to specific use cases need to be considered, as they have practical implications on society that can impact the public's perceptions and trust. Accordingly, this can enhance the development of ethical, effective, and accepted AI systems within the public sector, ultimately contributing to a more responsible and inclusive digital society.

6.3 Limitations and Future Research

This research is also subject to several limitations, which are important to note. First, a single case study of the Netherlands was used, which impacts the generalisability of the results. A unique context of advanced AI development, declining public trust and the legacy of past ethical failings is found in the Netherlands, which may impact Dutch citizens' perceptions on AI in a manner that may not apply in other contexts. Nevertheless, focusing on a single case enabled a more nuanced understanding of the local context and provided the opportunity for novel aspects to emerge. Insights from the Netherlands can thus aid AI developers, implementers and policy makers in other countries with the deployment of AI, whilst also ensuring that relevant concerns and societal needs are considered.

The second limitation relates to data collection. The method of convenience sampling in this research hindered opportunities of obtaining a representative sample from the Dutch population. In particular, the sample consists predominantly of highly educated citizens and clustered around two main age groups (i.e., 20–35 and >55), indicating a level of familiarity with technology and knowledge about risks. These elements impact the perceptions of participants, which might not be representative of broader society's views. Despite this, this research provides an initial exploratory view on Dutch citizens perceptions of the ethical implications of AI use in the public sector, which is a perspective thus far underdeveloped in research.

Third, the results and subsequent discussion of the ethical tensions and leading values may not be generalisable beyond the specific contexts they are applied to. Participants may have personal experiences with the cases discussed under the ethical tensions, impacting their responses. This can limit the generalisability of the values deemed leading when ethical tensions emerge, as such perceptions are impacted by the peculiarities of the specific case and can lead to different results based on the context. Nevertheless, the initial exploration of leading values in ethical tensions aid in underpinning salient elements and understanding citizens' experiences.

Beyond these limitations, this research extends valuable insights on citizens' perceptions within the domain of AI and ethics. There are multiple aspects of this research that can be expanded on to further knowledge and insights in this domain. One aspect includes promoting a deeper understanding of the specific ethical AI tensions. Each of the tensions identified — efficiency vs. privacy, accuracy vs. fairness, personalisation vs. solidarity, and automation vs. dignity—can be further explored. Research can delve into the nuances of these tensions, how they manifest in different contexts, and potential strategies for managing them.

Regarding the need to balance different values in AI design and implementation, future studies can investigate potential mechanisms or frameworks to achieve this balance, such as models for incorporating public input or decision-making strategies when navigating ethical trade-offs. The impact of public participation in AI design and implementation, including its effects on policy outcomes, public trust, and democratic legitimacy can also be assessed. This assessment can help with discerning the effects of incorporating diverse perspectives on the social acceptance of AI use. Moreover, given the low level of trust in the Dutch government, future studies can explore how the adoption of ethical AI principles in the public sector affects this trust. A longitudinal study is fit for this purpose and can map the interlinkages between the implementation of principles, changing citizens' perceptions of the ethical implications of AI use and the level of trust in the government, particularly in relation to advancing citizens' best interests.

These suggestions for future research should not only be considered relevant for the Netherlands, but offer a more expansive scope. For instance, this research can be replicated with a larger sample of participants, alongside utilising a cross-unit analysis of different countries to enhance generalisability. Divergences between citizens' perceptions in varying contexts can thus be analysed, providing valuable insights on the impact of local contexts on ethical understandings.

By delving deeper into these areas, insights for the ongoing development and implementation of ethical AI in the Dutch public sector and beyond can be expanded. The emphasis on and the future incorporation of citizens' perspectives can potentially diminish current negative perceptions associated with the ethical implications of AI use, heightening citizens' and broader society's opportunities to reap the expansive benefits that AI has to offer.

References

Abedin, B. 2021. "Managing the Tension Between Opposing Effects of Explainability of Artificial Intelligence: A Contingency Theory Perspective," *Internet Research* (32:2), pp. 425–453. (https://doi.org/10.1108/INTR-05-2020-0300).

Adams, S., Arel, I., Bach, J., Coop, R., Furlan, R., Goertzel, B., Hall, J. S., Samsonovich, A., Scheutz, M., Schlesinger, M., Shapiro, S. C., and Sowa, J. 2012. "Mapping the Landscape of Human-Level Artificial General Intelligence," *AI Magazine* (33:1), pp. 25-42. (https://doi.org/10.1609/aimag.v33i1.2322).

Ahn, M. J., and Chen, Y.-C. 2022. "Digital Transformation toward AI-Augmented Public Administration: The Perception of Government Employees and the Willingness to Use AI in Government," *Government Information Quarterly* (39:2), p. 101664. (https://doi.org/10.1016/j.giq.2021.101664).

Akter, S., McCarthy, G., Sajib, S., Michael, K., Dwivedi, Y. K., D'Ambra, J., and Shen, K. N. 2021. "Algorithmic Bias in Data-Driven Innovation in the Age of AI," *International Journal of Information Management* (60), p. 102387. (https://doi.org/10.1016/j.ijinfomgt.2021.102387).

Aldane, J. 2023. "Press OK for AI: Are Governments Ready to Automate?," *Global Government Forum*, February 12. (https://www.globalgovernmentforum.com/press-ok-for-ai-are-governments-ready-to-automate/, accessed May 5, 2023).

Almada, M. 2019. "Human Intervention in Automated Decision-Making: Toward the Construction of Contestable Systems," in *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Law*, pp. 2–11. (https://doi.org/10.1145/3322640.3326699).

Amnesty. 2021. "Dutch childcare benefit scandal an urgent wake-up call to ban racist algorithms," *Amnesty International*, October 25. (https://www.amnesty.org/en/latest/news/2021/10/xenophobic-machines-dutch-child-benefit-scandal/, accessed March 10, 2023).

Anderson, M., and Anderson, S. L. 2007. "Machine Ethics: Creating an Ethical Intelligent Agent," *AI Magazine* (28:4), pp. 15–15. (https://doi.org/10.1609/aimag.v28i4.2065).

Aoki, N. 2020. "An Experimental Study of Public Trust in AI Chatbots in the Public Sector," *Government Information Quarterly* (37:4), p. 101490. (https://doi.org/10.1016/j.giq.2020.101490).

Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., and Herrera, F. 2020. "Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI," *Information Fusion* (58), pp. 82–115. (https://doi.org/10.1016/j.inffus.2019.12.012).

Ashok, M., Madan, R., Joha, A., and Sivarajah, U. 2022. "Ethical Framework for Artificial Intelligence and Digital Technologies," *International Journal of Information Management* (62), p. 102433. (https://doi.org/10.1016/j.ijinfomgt.2021.102433).

Asilomar AI Principles. 2017. "AI Principles," *Future of Life Institute*, September 18. (https://futureoflife.org/open-letter/ai-principles/, accessed February 11, 2023).

Asser Institute. 2021. "[Call for Papers] Artificial Intelligence: The New Frontier of Business and Human Rights," *Asser Institute: Centre for International and European Law*, May 19. (https://www.asser.nl/about-the-asser-institute/news/call-for-papers-artificial-intelligence-the-new-frontier-of-business-and-human-rights/, accessed March 2, 2023).

Asveld, L., and Roeser, S. 2009. *The Ethics of Technological Risk*, London: Routledge.

Bao, L., Krause, N. M., Calice, M. N., Scheufele, D. A., Wirz, C. D., Brossard, D., Newman, T. P., and Xenos, M. A. 2022. "Whose AI? How Different Publics Think about AI and Its Social Impacts," *Computers in Human Behavior* (130), p. 107182. (https://doi.org/10.1016/j.chb.2022.107182).

Bataller, C., and Harris, J. 2016. "Turning Artificial Intelligence into Business Value. Today," *Accenture Institute for High Performance*, Dublin: Accenture. (https://files.stample.co/browserUpload/fc3be0d1-906c-4db4-b572-649edf4c73ac, accessed May 30, 2023).

Batra, G., Queirolo, A., and Santhanam, N. 2018. "Artificial Intelligence: The Time to Act Is Now| McKinsey," *McKinsey & Company*, January 8. (https://www.mckinsey.com/industries/industrials-and-electronics/our insights/artificial-intelligence-the-time-to-act-is-now, accessed January 20, 2023).

Beauchamp, T. L., and Childress, J. F. 2001. *Principles of biomedical ethics*, Oxford: Oxford University Press.

Beck, L. C., Trombetta, W. L., and Share, S. 1986. "Using focus group sessions before decisions are made," *North Carolina medical journal* (47:2), pp. 73–74. (https://pubmed.ncbi.nlm.nih.gov/3457274/)

Bellamy, R. K. E., Dey, K., Hind, M., Hoffman, S. C., Houde, S., Kannan, K., Lohia, P., Martino, J., Mehta, S., Mojsilović, A., Nagar, S., Ramamurthy, K. N., Richards, J., Saha, D., Sattigeri, P., Singh, M., Varshney, K. R., and Zhang, Y. 2019. "AI Fairness 360: An Extensible Toolkit for Detecting and Mitigating Algorithmic Bias," *IBM Journal of Research and Development* (63:4/5), pp. 4:1–4:15. (https://doi.org/10.1147/JRD.2019.2942287).

Benke, K., and Benke, G. 2018. "Artificial Intelligence and Big Data in Public Health," *International Journal of Environmental Research and Public Health* (15:12), p. 2796. (https://doi.org/10.3390/ijerph15122796).

Berendt, B. 2019. "AI for the Common Good?! Pitfalls, Challenges, and Ethics Pen-Testing," *Paladyn, Journal of Behavioral Robotics* (10:1), pp. 44–65. (https://doi.org/10.1515/pjbr-2019-0004).

Bingham, A. J., and Witkowsky, P. 2022. "Deductive and inductive approaches to qualitative data analysis," in *Analyzing and interpreting qualitative data: After the interview*, C. Vanover, P. Mihas, and J. Saldaña (eds.), California: Sage Publications, pp. 133-146.

Boztas, S. 2021. "KvK Made More than €3.3m in Last Years of 'unlawful' Data Sale," *DutchNews*, June 21. (https://www.dutchnews.nl/news/2021/06/kvks-made-more-than-e3-3m-in-last-years-of-illegal-data-sale/, accessed May 11, 2023).

Boyd, M., and Wilson, N. 2017. “Rapid Developments in Artificial Intelligence: How Might the New Zealand Government Respond?,” *Policy Quarterly* (13:04), pp. 36–43. (https://doi.org/10.26686/pq.v13i4.4619).

Braun, V., and Clarke, V. 2006. “Using Thematic Analysis in Psychology,” *Qualitative Research in Psychology* (3:2), pp. 77–101. (https://doi.org/10.1191/1478088706qp063oa).

Braun, V., and Clarke, V. 2012. “Thematic Analysis,” in *APA Handbook of Research Methods in Psychology, Vol 2: Research Designs: Quantitative, Qualitative, Neuropsychological, and Biological*, H. A. Cooper (ed.), Washington: American Psychological Association, pp. 57–71.

Braun, V., and Clarke, V. 2019. “Reflecting on Reflexive Thematic Analysis,” *Qualitative Research in Sport, Exercise and Health* (11:4), pp. 589–597. (https://doi.org/10.1080/2159676X.2019.1628806).

Breen, R. L. 2006. “A Practical Guide to Focus-Group Research,” *Journal of Geography in Higher Education* (30:3), pp. 463–475. (https://doi.org/10.1080/03098260600927575).

Brown, M. 2020. “A landmark court ruling could transform how governments use A.I.,” *Innovation Inverse*, February 7. (https://www.inverse.com/innovation/a-landmark-court-ruling-could-transform-how-governments-use-ai, accessed March 13, 2023).

Buhmann, A., Paßmann, J., and Fieseler, C. 2019. “Managing Algorithmic Accountability: Balancing Reputational Concerns, Engagement Strategies, and the Potential of Rational Discourse,” *Journal of Business Ethics* (163), pp. 265–280. (https://doi.org/10.1007/s10551-019-04226-4).

Buranyi, S. 2017. “Rise of the Racist Robots – How AI Is Learning All Our Worst Impulses,” *The Guardian*, August 8. (https://www.theguardian.com/inequality/2017/aug/08/rise-of-the-racist-robots-how-ai-is-learning-all-our-worst-impulses, accessed January 28, 2023).

Burrows, D., and Kendall, S. 1997. “Focus groups: what are they and how can they be used in nursing and health care research?,” *Social Sciences in Health* (3), pp. 244-253.

Busuioc, M. 2022. “AI Algorithmic Oversight: New Frontiers in Regulation,” in *Handbook of Regulatory Authorities*, M. Maggetti., F. D. Mascio., and A. Natalini (eds.), Cheltenham: Edward Elgar Publishing, pp. 470–486.

Byrne, D. 2022. “A Worked Example of Braun and Clarke’s Approach to Reflexive Thematic Analysis,” *Quality & Quantity* (56:3), pp. 1391–1412. (https://doi.org/10.1007/s11135-021-01182-y).

Calo, M. R. 2011. “Peeping Hals,” *Artificial Intelligence* (175:5), pp. 940–941. (https://doi.org/10.1016/j.artint.2010.11.025).

Cam, A., Chui, M., and Hall, B. 2019. “Global AI Survey: AI proves its worth, but few scale impact,” *McKinsey Analytics*, New York: McKinsey & Company. (http://dln.jaipuria.ac.in:8080/jspui/bitstream/123456789/1323/1/Global-AI-Survey-AI- proves-its-worth-but-few-scale-impact.pdf, accessed May 30, 2023).

Cath, C. 2018. “Governing Artificial Intelligence: Ethical, Legal and Technical Opportunities and Challenges,” *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* (376:2133), p. 20180080. (https://doi.org/10.1098/rsta.2018.0080).

[CBS] Centraal Bureau voor de Statistiek. 2020. “The Netherlands ranks among the EU top in digital skills,” *Statistics Netherlands*, February 14. (https://www.cbs.nl/en-gb/news/2020/07/the-netherlands-ranks-among-the-eu-top-in-digital-skills, accessed May 16, 2023).

Čerka, P., Grigienė, J., and Sirbikytė, G. 2017. “Is It Possible to Grant Legal Personality to Artificial Intelligence Software Systems?,” *Computer Law & Security Review* (33:5), pp. 685–699. (https://doi.org/10.1016/j.clsr.2017.03.022).

Chaves, A. P., and Gerosa, M. A. 2018. “Single or multiple conversational agents? An interactional coherence comparison,” in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pp. 1-13. (https://doi.org/10.1145/3173574.3173765).

Chen, L., Chen, P., and Lin, Z. 2020. “Artificial Intelligence in Education: A Review,” *IEEE Access* (8), pp. 75264–75278. (https://doi.org/10.1109/ACCESS.2020.2988510).

Chen, T., Guo, W., Gao, X., and Liang, Z. 2021. “AI-Based Self-Service Technology in Public Service Delivery: User Experience and Influencing Factors,” *Government Information Quarterly* (38:4), p. 101520. (https://doi.org/10.1016/j.giq.2020.101520).

Chun, A. H. W. 2008. “An AI Framework for the Automatic Assessment of E-Government Forms,” *AI Magazine* (29:1), pp. 52–52. (https://doi.org/10.1609/aimag.v29i1.2086).

Clouser, K. D., and Gert, B. 1990. “A Critique of Principlism,” *The Journal of Medicine and Philosophy: A Forum for Bioethics and Philosophy of Medicine* (15:2), pp. 219–236. (https://doi.org/10.1093/jmp/15.2.219).

Coeckelbergh, M. 2020. *AI Ethics*, Cambridge: MIT Press.

Cramer, B. W. 2008. “The Human Right to Information, the Environment and Information About the Environment: From the Universal Declaration to the Aarhus Convention,” *Communication Law and Policy* (14:1), pp. 73–103. (https://doi.org/10.1080/10811680802577707).

Crawford, K., and Calo, R. 2016. “There Is a Blind Spot in AI Research,” *Nature* (538:7625), pp. 311–313. (https://doi.org/10.1038/538311a).

Dagevos, J., de Voogd-Hamelink, M., and Damen, R. 2022. “Gevestigd, maar niet thuis: Eerste bevindingen uit de Survey integratie migranten,” *The Netherlands Institute for Social Research*, The Hague: Sociaal en Cultureel Planbureau. (https://www.scp.nl/publicaties/publicaties/2022/10/11/gevestigd-maar-niet-thuis.-eerste-bevindingen-uit-de-survey-integratie-migranten-sim2020, accessed May 20, 2023).

de Bruin, B., and Floridi, L. 2017. “The Ethics of Cloud Computing,” *Science and Engineering Ethics* (23:1), pp. 21–39. (https://doi.org/10.1007/s11948-016-9759-0).

de Sousa, W. G., Melo, E. R. P., Bermejo, P. H. D. S., Farias, R. A. S., and Gomes, A. O. 2019. “How and Where Is Artificial Intelligence in the Public Sector Going? A Literature Review and Research Agenda,” *Government Information Quarterly* (36:4), p. 101392. (https://doi.org/10.1016/j.giq.2019.07.004).

de Voogd, E. 2019. “European cities’ first steps with chatbots: The Next Step in Digital Customer Service?,” *Interreg North Sea Region*, Groningen: European Regional Development Fund. (https://northsearegion.eu/media/10487/gepubliceerd-report-chatbots.pdf, accessed May 30, 2023).

Denning, P. J., and Rosenbloom, P. S. 2009. “The Profession of IT Computing: The Fourth Great Domain of Science,” *Communications of the ACM* (52:9), pp. 27–29. (https://doi.org/10.1145/1562164.1562176).

Desai, J. N., Pandian, S., and Vij, R. K. 2021. “Big Data Analytics in Upstream Oil and Gas Industries for Sustainable Exploration and Development: A Review,” *Environmental Technology & Innovation* (21), p. 101186. (https://doi.org/10.1016/j.eti.2020.101186).

[DESI] Digital Economy and Society Index. 2022. “Digital Economy and Society Index (DESI) 2022: The Netherlands,” *European Commission*, Brussels: European Commission. (https://digital-strategy.ec.europa.eu/en/library/digital-economy-and-society-index-desi-2022, accessed May 30, 2023).

Dietvorst, B. J., Simmons, J. P., and Massey, C. 2015. “Algorithm Aversion: People Erroneously Avoid Algorithms after Seeing Them Err,” *Journal of Experimental Psychology: General* (144), pp. 114–126. (https://doi.org/10.1037/xge0000033).

Easton, J. 2018. “Where to Draw the Line? Is Efficiency Encroaching on a Fair Justice System?,” *The Political Quarterly* (89:2), pp. 246–253. (https://doi.org/10.1111/1467-923X.12487).

[EC] European Commission. 2018. “Statement on artificial intelligence, robotics and 'autonomous' systems,” *European Group on Ethics in Science and New Technologies*, Brussels: European Commission. (https://doi.org/10.2777/531856).

[EC] European Commission. 2019a. “A Definition of AI: Main Capabilities and Scientific Disciplines,” *High Level Expert Group on Artificial Intelligence*, Brussels: European Commission. (https://digital-strategy.ec.europa.eu/en/library/definition-artificial-intelligence-main-capabilities-and-scientific-disciplines, accessed May 30, 2023).

[EC] European Commission. 2019b. “Ethics Guidelines for Trustworthy AI,” *High Level Expert Group on Artificial Intelligence*, Brussels: European Commission. (https://www.europarl.europa.eu/cmsdata/196377/AI%20HLEG_Ethics%20Guidelines%20for%20Trustworthy%20AI.pdf, accessed May 30, 2023).

[ECNL] European Center for Not-for-Profit Law. 2022. “Netherlands Sets Precedent for Human Rights Safeguards in Use of AI,” *ECNL*, April 12. (https://ecnl.org/news/netherlands-sets-precedent-human-rights-safeguards-use-ai, accessed January 25, 2023).

Farrell, E. H., Whistance, R. N., Phillips, K., Morgan, B., Savage, K., Lewis, V., Kelly, M., Blazeby, J. M., Kinnersley, P., and Edwards, A. 2014. "Systematic Review and Meta-Analysis of Audio-Visual Information Aids for Informed Consent for Invasive Healthcare Procedures in Clinical Practice," *Patient Education and Counseling* (94:1), pp. 20–32. (https://doi.org/10.1016/j.pec.2013.08.019).

Floridi, L. 2013. *The Ethics of Information*, Oxford: Oxford University Press.

Floridi, L. 2016. "On human dignity as a foundation for the right to privacy," *Philosophy & Technology* (29), pp. 307-312. (https://doi.org/10.1007/s13347-016-0220-8)

Floridi, L. 2018. "Soft Ethics, the Governance of the Digital and the General Data Protection Regulation," *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* (376:2133), p. 20180081. (https://doi.org/10.1098/rsta.2018.0081).

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., Schafer, B., Valcke, P., and Vayena, E. 2018. "AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations," *Minds and Machines* (28:4), pp. 689–707. (https://doi.org/10.1007/s11023-018-9482-5).

Flyvbjerg, B. 2006. "Five Misunderstandings About Case-Study Research," *Qualitative Inquiry* (12:2), pp. 219–245. (https://doi.org/10.1177/1077800405284363).

Fonteyn, M. E., Vettese, M., Lancaster, D. R., and Bauer-Wu, S. 2008. "Developing a Codebook to Guide Content Analysis of Expressive Writing Transcripts," *Applied Nursing Research* (21:3), pp. 165–168. (https://doi.org/10.1016/j.apnr.2006.08.005).

Franzke, A. S., Muis, I., and Schäfer, M. T. 2021. "Data Ethics Decision Aid (DEDA): A Dialogical Framework for Ethical Inquiry of AI and Data Projects in the Netherlands," *Ethics and Information Technology* (23:3), pp. 551–567. (https://doi.org/10.1007/s10676-020-09577-5).

Friese, S. 2019. *Qualitative Data Analysis with ATLAS.Ti*, California: Sage Publications.

Gartner. 2018. "Gartner Says Nearly Half of CIOs Are Planning to Deploy Artificial Intelligence," *Gartner Inc.*, February 13. (https://www.gartner.com/en/newsroom/press-releases/2018-02-13-gartner-says-nearly-half-of-cios-are-planning-to-deploy-artificial-intelligence, accessed February 18, 2023).

Gerrish, C., and Morrison, L. 2020. "Can the Law Keep Up with the Growth of AI?," in *The LegalTech Book: The Legal Technology Handbook for Investors, Entrepreneurs and FinTech Visionaries*, S. A. Bhatti, S. Chishti, A. Datoo, and D. Indjic (eds.), New Jersey: John Wiley and Sons Inc., pp. 30–34.

Gesk, T. S., and Leyer, M. 2022. "Artificial intelligence in public services: When and why citizens accept its usage," *Government Information Quarterly* (39:3), pp. 101704. (https://doi.org/10.1016/j.giq.2022.101704)

Gillespie, T., Boczkowski, P. J., and Foot, K. A. 2014. *Media Technologies: Essays on Communication, Materiality, and Society*, Cambridge: MIT Press.

Greene, D., Hoffmann, A. L., and Stark, L. 2019. “Better, Nicer, Clearer, Fairer: A Critical Assessment of the Movement for Ethical Artificial Intelligence and Machine Learning,” in *Proceedings of the 52nd Hawaii International Conference on System Sciences*, pp. 2122–2131. (http://dx.doi.org/10.24251/HICSS.2019.258).

Grewal, D., Kroschke, M., Mende, M., Roggeveen, A. L., and Scott, M. L. 2020. “Frontline Cyborgs at Your Service: How Human Enhancement Technologies Affect Customer Experiences in Retail, Sales, and Service Settings,” *Journal of Interactive Marketing* (51), pp. 9–25. (https://doi.org/10.1016/j.intmar.2020.03.001).

Guba, E. G., and Lincoln, Y. S. 1994. “Competing Paradigms in Qualitative Research,” in *Handbook of Qualitative Research*, N. K. Denzin and Y. S. Lincoln (eds.), California: Sage Publications, pp. 105–117.

Gurr, T. R. 1971. *Why Men Rebel*, New Jersey: Princeton University Press.

Habermas, J. 1991. *Erläuterungen zur diskursethik*, Berlin: Suhrkamp Verlag.

Hagendorff, T. 2020. “The Ethics of AI Ethics: An Evaluation of Guidelines,” *Minds and Machines* (30:1), pp. 99–120. (https://doi.org/10.1007/s11023-020-09517-8).

Hao, K. 2020. “The UK Exam Debacle Reminds Us That Algorithms Can’t Fix Broken Systems,” *MIT Technology Review*, August 20. (https://www.technologyreview.com/2020/08/20/1007502/uk-exam-algorithm-cant-fix-broken-system/, accessed January 20, 2023).

Hashimi, A. 2019. “AI Ethics: The Next Big Thing in Government – Anticipating the Impacts of AI Ethics within the Public Sector,” *World Government Summit and Deloitte*, Dubai: Deloitte. (https://www2.deloitte.com/content/dam/Deloitte/xe/Documents/AboutDeloitte/WGS%20report%20l%20AI%20Ethics.pdf, accessed May 30, 2023).

Heaven, W. D. 2023. “ChatGPT Is Going to Change Education, Not Destroy It,” *MIT Technology Review*, April 6. (https://www.technologyreview.com/2023/04/06/1071059/chatgpt-change-not-destroy-education-openai/, accessed May 16, 2023).

Héder, M. 2020. “A Criticism of AI Ethics Guidelines,” *Információs Társadalom* (20:4), p. 57–73. (https://doi.org/10.22503/inftars.XX.2020.4.5).

Heikkilä, M. 2022. “Dutch Scandal Serves as a Warning for Europe over Risks of Using Algorithms,” *Politico*, March 29. (https://www.politico.eu/article/dutch-scandal-serves-as-a-warning-for-europe-over-risks-of-using-algorithms/, accessed May 15, 2023).

Heilinger, J.-C. 2022. “The Ethics of AI Ethics. A Constructive Critique,” *Philosophy & Technology* (35:3), p. 61. (https://doi.org/10.1007/s13347-022-00557-9).

Hickok, M. 2021. “Lessons Learned from AI Ethics Principles for Future Actions,” *AI and Ethics* (1:1), pp. 41–47. (https://doi.org/10.1007/s43681-020-00008-1).

Hildebrandt, M. 2016. “Law as Information in the Era of Data-Driven Agency,” *The Modern Law Review* (79:1), pp. 1–30. (https://doi.org/10.1111/1468-2230.12165).

IEEE Initiative on Ethics of Autonomous and Intelligent Systems. 2017. "Ethically Aligned Design: A Vision for Prioritising Human Well-being with Autonomous and Intelligent Systems," *IEEE Standards Association*, New Jersey: IEEE. (https://standards.ieee.org/wp-content/uploads/import/documents/other/ead_v2.pdf, accessed May 30, 2023).

Ingrams, A., Kaufmann, W., and Jacobs, D. 2022. "In AI We Trust? Citizen Perceptions of AI in Government Decision Making," *Policy & Internet* (14:2), pp. 390–409. (https://doi.org/10.1002/poi3.276).

Jarrahi, M. H. 2018. "Artificial Intelligence and the Future of Work: Human-AI Symbiosis in Organizational Decision Making," *Business Horizons* (61:4), pp. 577–586. (https://doi.org/10.1016/j.bushor.2018.03.007).

Jobin, A., Ienca, M., and Vayena, E. 2019. "The Global Landscape of AI Ethics Guidelines," *Nature Machine Intelligence* (1:9), pp. 389–399. (https://doi.org/10.1038/s42256-019-0088-2).

Johnson, A. 1996. "'It's Good to Talk': The Focus Group and the Sociological Imagination," *The Sociological Review* (44:3), pp. 517–538. (https://doi.org/10.1111/j.1467-954X.1996.tb00435.x).

Karliuk, M. 2022. "Proportionality principle for the ethics of artificial intelligence," *AI Ethics*, pp. 1–6. (https://doi.org/10.1007/s43681-022-00220-1).

Kessler, G. 2018. "Technology and the Future of Language Teaching," *Foreign Language Annals* (51:1), pp. 205–218. (https://doi.org/10.1111/flan.12318).

Kieslich, K., Keller, B., and Starke, C. 2022. "Artificial Intelligence Ethics by Design. Evaluating Public Perception on the Importance of Ethical Design Principles of Artificial Intelligence," *Big Data & Society* (9:1), p. 205395172210929. (https://doi.org/10.1177/20539517221092956).

Kitchin, R. 2017. "Thinking Critically about and Researching Algorithms," *Information, Communication & Society* (20:1), pp. 14–29. (https://doi.org/10.1080/1369118X.2016.1154087).

Kitzinger, J. 1994. "The Methodology of Focus Groups: The Importance of Interaction between Research Participants," *Sociology of Health & Illness* (16:1), pp. 103–121. (https://doi.org/10.1111/1467-9566.ep11347023).

Knowles, B. and Hanson, V. L. 2018. "The wisdom of older technology (non) users," *Communications of the ACM* (61:3), pp. 72–77. (https://doi.org/10.1145/3179995).

Kokkinidis, T. 2022. "Artificial Intelligence Will 'Likely' Destroy Humans, Researchers Say," *GreekReporter*, September 16. (https://greekreporter.com/2022/09/16/artificial-intelligence-annihilate-humankind/, accessed January 20, 2023).

König, P. D. 2022. "Citizen Conceptions of Democracy and Support for Artificial Intelligence in Government and Politics," *European Journal of Political Research*, pp. 1–21. (https://doi.org/10.1111/1475-6765.12570).

Krämer, C. 2020. “Can Robots Have Dignity?,” in *Artificial Intelligence: Reflections in Philosophy, Theology, and the Social Sciences*, A. Rosenthal-Von Der Putten, and B. P. Gocke (eds.), Leiden: Brill U Mentis, pp. 241-253.

Krueger, R. A. 1988. *Focus Groups: A Practical Guide for Applied Research*, California: Sage Publications.

Kuziemski, M., and Misuraca, G. 2020. “AI Governance in the Public Sector: Three Tales from the Frontiers of Automated Decision-Making in Democratic Settings,” *Telecommunications Policy* (44:6), p. 101976. (https://doi.org/10.1016/j.telpol.2020.101976).

Latour, B. 2007. *Reassembling the Social: An Introduction to Actor-Network-Theory*, Oxford: Oxford University Press.

Lauer, D. 2021. “You Cannot Have AI Ethics without Ethics,” *AI and Ethics* (1:1), pp. 21–25. (https://doi.org/10.1007/s43681-020-00013-4).

Lee, M. K. 2018. “Understanding Perception of Algorithmic Decisions: Fairness, Trust, and Emotion in Response to Algorithmic Management,” *Big Data & Society* (5:1), p. 205395171875668. (https://doi.org/10.1177/2053951718756684).

Lee, S., Nah, S., Chung, D. S., and Kim, J. 2020. “Predicting AI News Credibility: Communicative or Social Capital or Both?,” *Communication Studies* (71:3), pp. 428–447. (https://doi.org/10.1080/10510974.2020.1779769).

Lee, T. D., Lee-Geiller, S. and Lee, B. K. 2020. “Are pictures worth a thousand words? The effect of information presentation type on citizen perceptions of government websites,” *Government Information Quarterly* (37:3), p. 101482. (https://doi.org/10.1016/j.giq.2020.101482).

Legg, S., and Hutter, M. 2006. “A Formal Measure of Machine Intelligence,” in *Proceedings of the 15th Annual Machine Learning Conference of {B}elgium and The Netherlands*, pp. 73–80. (https://doi.org/10.48550/arXiv.cs/0605024).

Leikas, J., Koivisto, R., and Gotcheva, N. 2019. “Ethical Framework for Designing Autonomous Intelligent Systems,” *Journal of Open Innovation: Technology, Market, and Complexity* (5:1), p. 18. (https://doi.org/10.3390/joitmc5010018).

Liu, K. 2000. *Semiotics in Information Systems Engineering*, Cambridge: Cambridge University Press.

Loohuis, K. 2022. “Netherlands Coalition Aims to Demystify Artificial Intelligence,” *ComputerWeekly*, November 16. (https://www.computerweekly.com/news/252527232/Netherlands-coalition-aims-to-demystify-artificial-intelligence, accessed May 5, 2023).

Luengo-Oroz, M. 2019. “Solidarity should be a core ethical principle of AI,” *Nature Machine Intelligence* (1:11), pp. 494. (https://doi.org/10.1038/s42256-019-0115-3).

Luomo-aho, V., and Canel, M-J. 2020. *The Handbook of Public Sector Communication*, New Jersey: John Wiley & Sons Inc.

Maarse, J. A. M., and Jeurissen, P. P. 2016. "The Policy and Politics of the 2015 Long-Term Care Reform in the Netherlands," *Health Policy* (120:3), pp. 241–245. (https://doi.org/10.1016/j.healthpol.2016.01.014).

Maragno, G., Tangi, L., Gastaldi, L., and Benedetti, M. 2022. "The Spread of Artificial Intelligence in the Public Sector: A Worldwide Overview," in *Proceedings of the 14th International Conference on Theory and Practice of Electronic Governance*, pp. 1–9. (https://doi.org/10.1145/3494193.3494194).

Marcinkowski, F., Kieslich, K., Starke, C., and Lünich, M. 2020. "Implications of AI (Un-)Fairness in Higher Education Admissions: The Effects of Perceived AI (Un-)Fairness on Exit, Voice and Organizational Reputation," in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pp. 122–130. (https://doi.org/10.1145/3351095.3372867).

Margetts, H. 2022. "Rethinking AI for Good Governance," *Daedalus* (151:2), pp. 360–371. (https://doi.org/10.1162/daed_a_01922).

McCulloch, W. S., and Pitts, W. 1943. "A Logical Calculus of the Ideas Immanent in Nervous Activity," *The Bulletin of Mathematical Biophysics* (5:4), pp. 115–133. (https://doi.org/10.1007/BF02478259).

McDonald, H. 2020. "Home Office to Scrap 'racist Algorithm' for UK Visa Applicants," *The Guardian*, August 4. (https://www.theguardian.com/uk-news/2020/aug/04/home-office-to-scrap-racist-algorithm-for-uk-visa-applicants, accessed February 12, 2023).

McIntyre, A. 1988. *Whose Justice? Which Rationality?*, Indiana: University of Notre Dame Press.

McKnight, D. H., Choudhury, V., and Kacmar, C. 2002. "Developing and Validating Trust Measures for E-Commerce: An Integrative Typology," *Information Systems Research* (13:3), pp. 334–359. (https://doi.org/10.1287/isre.13.3.334.81).

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., and Galstyan, A. 2021. "A Survey on Bias and Fairness in Machine Learning," *Communications of the Association for Computing Machinery* (54:6), pp. 1–35. (https://doi.org/10.1145/3457607).

Meijer, A., and Wessels, M. 2019. "Predictive Policing: Review of Benefits and Drawbacks," *International Journal of Public Administration* (42:12), pp. 1031–1039. (https://doi.org/10.1080/01900692.2019.1575664).

Metzinger, T. 2019. "EU Guidelines: Ethics Washing Made in Europe," *Tagesspiegel*, April 8. (https://www.tagesspiegel.de/politik/ethics-washing-made-in-europe-5937028.html, accessed January 23, 2023).

Mikhaylov, S. J., Esteve, M., and Campion, A. 2018. "Artificial Intelligence for the Public Sector: Opportunities and Challenges of Cross-Sector Collaboration," *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* (376:2128), p. 20170357. (https://doi.org/10.1098/rsta.2017.0357).

Miller, S. M., and Keiser, L. R. 2021. "Representative Bureaucracy and Attitudes Toward Automated Decision Making," *Journal of Public Administration Research and Theory* (31:1), pp. 150–165. (https://doi.org/10.1093/jopart/muaa019).

Ministry of Interior and Kingdom Relations. 2019. “Beleidsbrief AI, publieke waarden en mensenrechten,” *Digital Overheid*, Den Haag: Rijksoverheid. (https://www.digitaleoverheid.nl/wp-content/uploads/sites/8/2020/01/Dutch-policy-brief-on-AI-public-values-and-fundamental-rights_DEF-T.pdf, accessed May 30, 2023).

Misuraca, G., and van Noordt, C. 2020. “AI Watch - Artificial Intelligence in Public Services: Overview of the Use and Impact of AI in Public Services in the EU,” *JRC Research Reports*, Seville: Joint Research Centre. (https://publications.jrc.ec.europa.eu/repository/handle/JRC120399, accessed May 30, 2023).

Misuraca, G., van Noordt, C., and Boukli, A. 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*, pp. 90–99. (https://doi.org/10.1145/3428502.3428513).

Mittelstadt, B. 2019. “Principles Alone Cannot Guarantee Ethical AI,” *Nature Machine Intelligence* (1:11), pp. 501–507. (https://doi.org/10.1038/s42256-019-0114-4).

Moingeon, P., Kuenemann, M., and Guedj, M. 2022. “Artificial Intelligence-Enhanced Drug Design and Development: Toward a Computational Precision Medicine,” *Drug Discovery Today* (27:1), pp. 215–222. (https://doi.org/10.1016/j.drudis.2021.09.006).

Mökander, J., and Floridi, L. 2021. “Ethics-Based Auditing to Develop Trustworthy AI,” *Minds and Machines* (31:2), pp. 323–327. (https://doi.org/10.1007/s11023-021-09557-8).

Montreal Declaration. 2017. “Montreal Declaration Activity Report 2018-2022”, *Université de Montréal*, Montréal: Université de Montréal et du monde. (https://www.montrealdeclarationresponsibleai.com/_files/ugd/ebc3a3_269e4424ff934998bc9fc0c1552e865b.pdf, accessed May 30, 2023).

Mooi, R. 2021. “Dutch Data Scandal Highlights Structural Problems around Privacy Compliance,” *IAPP News*, February 4. (https://iapp.org/news/a/dutch-data-scandal-highlights-structural-problems-around-privacy-compliance/, accessed May 9, 2023).

Moor, J. H. 2006. “The Nature, Importance, and Difficulty of Machine Ethics,” *IEEE Intelligent Systems* (21:4), pp. 18–21. (https://doi.org/10.1109/MIS.2006.80).

Morgan, D. L., Morgan, D. L., and Krueger, R. A. 1998. *The Focus Group Guidebook*, California: Sage Publications.

Morley, J., Floridi, L., Kinsey, L., and Elhalal, A. 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* (26:4), pp. 2141–2168. (https://doi.org/10.1007/s11948-019-00165-5).

Munn, L. 2022. “The Uselessness of AI Ethics,” *AI and Ethics*, pp. 1–9. (https://doi.org/10.1007/s43681-022-00209-w).

Nakao, Y., Stumpf, S., Ahmed, S., Naseer, A., and Strappelli, L. 2022. "Toward Involving End-Users in Interactive Human-in-the-Loop AI Fairness," *ACM Transactions on Interactive Intelligent Systems* (12:3), pp. 1–30. (https://doi.org/10.1145/3514258).

Nasseef, O. A., Baabdullah, A. M., Alalwan, A. A., Lal, B., and Dwivedi, Y. K. 2022. "Artificial Intelligence-Based Public Healthcare Systems: G2G Knowledge-Based Exchange to Enhance the Decision-Making Process," *Government Information Quarterly* (39:4), p. 101618. (https://doi.org/10.1016/j.giq.2021.101618).

Newell, S., and Marabelli, M. 2015. "Strategic Opportunities (and Challenges) of Algorithmic Decision-Making: A Call for Action on the Long-Term Societal Effects of 'Datification,'" *The Journal of Strategic Information Systems* (24:1), pp. 3–14. (https://doi.org/10.1016/j.jsis.2015.02.001).

NOS. 2022. "CBS: vier op de tien Nederlanders voelen zich eenzaam," *NOS News*, September 29. (https://nos.nl/artikel/2446395-cbs-vier-op-de-tien-nederlanders-voelen-zich-eenzaam, accessed May 5, 2023).

Ogden, C. K., and Richards, I. A. 1989. *The Meaning of Meaning: A Study of the Influence of Language upon Thought and of the Science of Symbolism*, San Diego: Harcourt Brace Jovanovich.

Ouchchy, L., Coin, A., and Dubljević, V., 2020. "AI in the headlines: the portrayal of the ethical issues of artificial intelligence in the media," *AI & Society* (35), pp. 927-936. (https://doi.org/10.1007/s00146-020-00965-5)

Pan, Y. 2016. "Heading toward Artificial Intelligence 2.0," *Engineering* (2:4), pp. 409–413. (https://doi.org/10.1016/J.ENG.2016.04.018).

Partnership on AI. 2018. "Partnership on AI Tenets," *Ethics Code Collection*. (http://ethicscodescollection.org/detail/5de67e84-93b4-4374-862d-f7344443b2ca, accessed March 12, 2023).

Patrick, E. A., and Fattu, J. 1986. *Artificial intelligence with statistical pattern recognition*, New Jersey: Englewood Cliffs.

Patton, M. Q. 1990. *Qualitative Evaluation and Research Methods*, California: Sage Publications.

Pfotenhauer, S., and Jasanoff, S. 2017. "Panacea or Diagnosis? Imaginaries of Innovation and the 'MIT Model' in Three Political Cultures," *Social Studies of Science* (47:6), pp. 783–810. (https://doi.org/10.1177/0306312717706110).

Plaut, T., Landis, S., and Trevor, J. 1993. "Focus groups and community mobilization: A case study from rural North Carolina," in *Successful focus groups: Advancing the state of the art*, D. L. Morgan (ed.), California: Sage Publications, pp. 202–221.

Popper, K. 1979. *Three Worlds*, Michigan: University of Michigan.

Powles, J. 2018. "The Seductive Diversion of 'Solving' Bias in Artificial Intelligence," *OneZero*, December 7. (https://onezero.medium.com/the-seductive-diversion-of-solving-bias-in-artificial-intelligence-890df5e5ef53, accessed March 7, 2023).

Radulov, N. 2019. "Artificial intelligence and security. Security 4.0", *Security & Future* (3:1), pp. 3–5. (https://doi.org/10.1007/s10462-020-09942-2).

Rathenau. 2021. “Research on artificial intelligence in the Netherlands,” *Rathenau Instituut*, September 14. (https://www.rathenau.nl/en/science-figures/research-artificial-intelligence-netherlands, accessed March 1, 2023).

Reid, K., Flowers, P., and Larkin, M. 2005. “Exploring Lived Experience,” *The Psychologist* (18:1), pp. 20–23.

Reuters. 2023. “Dutch Government to Face No-Confidence Vote after Election Loss,” *Reuters*, April 5. (https://www.reuters.com/world/europe/dutch-government-face-no-confidence-vote-after-election-loss-2023-04-05/, accessed May 9, 2023).

Rich, E., Knight, K., and Shivashankar, B. 2009. *Artificial Intelligence*, New Delhi: Tata McGraw Hill.

Riedl, M. O. 2019. “Human-Centered Artificial Intelligence and Machine Learning,” *Human Behavior and Emerging Technologies* (1:1), pp. 33–36. (https://doi.org/10.1002/hbe2.117).

Rijksoverheid. 2021. “The Dutch Digitalisation Strategy 2021,” *Nederland Digitaal*, Den Haag: Rijksoverheid. (https://www.nederlanddigitaal.nl/english/dutch-digitalisation-strategy-2.0, accessed May 30, 2023).

Roseman, G. H., and Stephenson, E. F. 2005. “The Effect of Voting Technology on Voter Turnout: Do Computers Scare the Elderly?,” *Public Choice* (123:1), pp. 39–47. (https://doi.org/10.1007/s11127-005-3993-3).

Rosemann, M., Becker, J., and Chasin, F. 2021. “City 5.0,” *Business & Information Systems Engineering* (63:1), pp. 71–77. (https://doi.org/10.1007/s12599-020-00674-9).

Rössler, B. 2004. *Privacies: Philosophical Evaluations*, California: Stanford University Press.

Russel, S., and Norvig, P. 2010. *Artificial Intelligence: A Modern Approach – Third Edition*, New Jersey: Englewood Cliffs.

Samoili, S., Lopez Cobo, M., Delipetrev, B., Martinez-Plumed, F., Gomez Gutierrez, E., and de Prato, G. 2021. “AI Watch: Defining Artificial Intelligence 2.0,” *Joint Research Centre*, Luxembourg City: Publications Office of the European Union. (https://dx.doi.org/10.2760/019901).

Schiff, D., Rakova, B., Ayesh, A., Fanti, A., and Lennon, M. 2020. “Principles to Practices for Responsible AI: Closing the Gap,” *arXiv preprint draft 2006.04707*. (https://doi.org/10.48550/ARXIV.2006.04707).

Schulte, I., Hart, D., and van der Vorst, R. 2004. “Issues Affecting the Acceptance of Hydrogen Fuel,” *International Journal of Hydrogen Energy* (29:7), pp. 677–685. (https://doi.org/10.1016/j.ijhydene.2003.09.006).

Serafimova, S. 2020. “Whose morality? Which rationality? Challenging artificial intelligence as a remedy for the lack of moral enhancement,” *Humanities and Social Sciences Communication* (7:119), pp. 1–10. (https://doi.org/10.1057/s41599-020-00614-8).

Smith, J. M. 1972. *Interviewing in market and social research,* Oxfordshire: Routledge and Kegan Paul Books.

Smith, B., and Welty, C. 2001. "FOIS Introduction: Ontology---towards a New Synthesis," in *Proceedings of the International Conference on Formal Ontology in Information Systems*, pp. .3–.9 (https://doi.org/10.1145/505168.505201).

Smith, J. A., Larkin, M., and Flowers, P. 2009. *Interpretative phenomenological analysis: theory, method and research,* California: Sage Publications.

Song, S., Chaudhuri, K., and Sarwate, A. D. 2013. "Stochastic Gradient Descent with Differentially Private Updates," in *2013 IEEE Global Conference on Signal and Information Processing*, pp. 245–248. (https://doi.org/10.1109/GlobalSIP.2013.6736861).

Starke, C., and Lünich, M. 2020. "Artificial Intelligence for Political Decision-Making in the European Union: Effects on Citizens' Perceptions of Input, Throughput, and Output Legitimacy," *Data & Policy* (2), p. e16. (https://doi.org/10.1017/dap.2020.19).

Statista. 2022. "Netherlands: Public Trust in the National Government 2022," *Statista*, September 22. (https://www.statista.com/statistics/587527/public-trust-in-the-national-government-in-the-netherlands/, accessed January 25, 2023).

Steffen, W., Richardson, K., Rockström, J., Cornell, S. E., Fetzer, I., Bennett, E. M., Biggs, R., Carpenter, S. R., de Vries, W., de Wit, C. A., Folke, C., Gerten, D., Heinke, J., Mace, G. M., Persson, L. M., Ramanathan, V., Reyers, B., and Sörlin, S. 2015. "Planetary Boundaries: Guiding Human Development on a Changing Planet," *Science* (347:6223), p. 1259855. (https://doi.org/10.1126/science.1259855).

Sukhadeve, A. 2021. "Council Post: Artificial Intelligence For Good: How AI Is Helping Humanity," *Forbes*, February 2. (https://www.forbes.com/sites/forbesbusinesscouncil/2021/02/09/artificial-intelligence-for-good-how-ai-is-helping-humanity/, accessed January 20, 2023).

Taebi, B. 2017. "Bridging the Gap between Social Acceptance and Ethical Acceptability," *Risk Analysis* (37:10), pp. 1817–1827. (https://doi.org/10.1111/risa.12734).

Tangi, L., Janssen, M., Benedetti, M., and Noci, G. 2020. "Barriers and Drivers of Digital Transformation in Public Organizations: Results from a Survey in the Netherlands," in *Electronic Government*, G. Viale Pereira, M. Janssen, H. Lee, I. Lindgren, M. P. Rodríguez Bolívar, H. J. Scholl, and A. Zuiderwijk (eds.), Cham: Springer International Publishing, pp. 42–56.

The People's Republic of China. 2017. "China Issues Guideline on Artificial Intelligence Development," *English Gov CN*, July 20. (http://english.www.gov.cn/policies/latest_releases/2017/07/20/content_281475742458322.htm, accessed January 19, 2023).

Thiebes, S., Lins, S., and Sunyaev, A. 2021. "Trustworthy Artificial Intelligence," *Electronic Markets* (31:2), pp. 447–464. (https://doi.org/10.1007/s12525-020-00441-4).

Toh, M. 2023. “300 Million Jobs Could Be Affected by Latest Wave of AI, Says Goldman Sachs,” *CNN News*, March 29. (https://www.cnn.com/2023/03/29/tech/chatgpt-ai-automation-jobs-impact-intl-hnk/index.html, accessed May 11, 2023).

Tortoise Media. 2022. “The Global AI Index,” *Tortoise Media Intelligence*, December 4. (https://www.tortoisemedia.com/intelligence/global-ai/, accessed March 6, 2023).

UK House of Lords. 2018. “House of Lords - AI in the UK: Ready, Willing and Able? - Artificial Intelligence Committee”, *Select Committee on Artificial Intelligence*, London: House of Lords. (https://publications.parliament.uk/pa/ld201719/ldselect/ldai/100/100.pdf, accessed May 30, 2023).

van Noordt, C., and Misuraca, G. 2019. “New Wine in Old Bottles: Chatbots in Government,” in *Electronic Participation*, P. Panagiotopoulos, N. Edelmann, O. Glassey, G. Misuraca, P. Parycek, T. Lampoltshammer, and B. Re (eds.), Cham: Springer International Publishing, pp. 49–59.

van Noort, W. 2015. “Hoe de slimme stad een dom idee kan worden,” *NRC Handelsblad*, October 17. (https://www.nrc.nl/nieuws/2015/10/17/de-slimme-stad-kan-een-dom-idee-worden-1546062-a289041, accessed March 11, 2023).

van Veenstra, A. F., Grommé, F. and Djafari, S. 2021. “The use of public sector data analytics in the Netherlands,” *Transforming Government: People, Process and Policy* (15:4), pp. 396–419. (https://doi.org/10.1108/TG-09-2019-0095).

Wang, W., and Siau, K. 2018. “Ethical and Moral Issues with AI,*”* in *The 24th Americas Conference on Information Systems*, pp. 1–5.

Waterman, D. A., and Newell, A. 1971. “Protocol Analysis as a Task for Artificial Intelligence,” *Artificial Intelligence* (2:3), pp. 285–318. (https://doi.org/10.1016/0004-3702(71)90014-2).

White, M. 2011. *Kantian Ethics and Economics: Autonomy, Dignity, and Character*, California: Princeton University Press.

Whittlestone, J., Nyrup, R., Alexandrova, A., and Cave, S. 2019. “The Role and Limits of Principles in AI Ethics: Towards a Focus on Tensions,” in *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 195–200. (https://doi.org/10.1145/3306618.3314289).

Wilkinson, S. 1998. “Focus Group Methodology: A Review,” *International Journal of Social Research Methodology* (1:3), pp. 181–203. (https://doi.org/10.1080/13645579.1998.10846874).

Willems, J., Schmid, M. J., Vanderelst, D., Vogel, D., and Ebinger, F. 2022. “AI-Driven Public Services and the Privacy Paradox: Do Citizens Really Care about Their Privacy?,” *Public Management Review*, pp. 1–19. (https://doi.org/10.1080/14719037.2022.2063934).

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

Yigitcanlar, T., Degirmenci, K., and Inkinen, T. 2022. "Drivers behind the Public Perception of Artificial Intelligence: Insights from Major Australian Cities," *AI & Society*, pp. 1–21. (https://doi.org/10.1007/s00146-022-01566-0).

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

Zhai, Y., Yan, J., Zhang, H., and Lu, W. 2020. "Tracing the Evolution of AI: Conceptualization of Artificial Intelligence in Mass Media Discourse," *Information Discovery and Delivery* (48:3), pp. 137–149. (https://doi.org/10.1108/IDD-01-2020-0007).

Zhai, X., Chu, X., Chai, C. S., Jong, M. S. Y., Istenic, A., Spector, M., Liu, J.-B., Yuan, J., and Li, Y. 2021. "A Review of Artificial Intelligence (AI) in Education from 2010 to 2020," *Complexity* (2021), p. e8812542. (https://doi.org/10.1155/2021/8812542).

Zhang, Q., Sun, H., Wu, X., and Zhong, H. 2019. "Edge Video Analytics for Public Safety: A Review," in *Proceedings of the IEEE* (107:8), pp. 1675–1696. (https://doi.org/10.1109/JPROC.2019.2925910).

Zhou, J., Chen, F., and Holzinger, A. 2022. "Towards Explainability for AI Fairness," in *AI - Beyond Explainable AI: International Workshop, Held in Conjunction with ICML 2020*, A. Holzinger, R. Goebel, R. Fong, T. Moon, K.-R. Müller, and W. Samek (eds.), Cham: Springer International Publishing, pp. 375–386.

Zhu, J., Hou, R., Wang, X., Wang, W., Cao, J., Zhao, B., Wang, Z., Zhang, Y., Ying, J., Zhang, L., and Meng, D. 2020. "Enabling Rack-scale Confidential Computing using Heterogeneous Trusted Execution Environment," in *Proceedings of the 2020 IEEE Symposium on Security and Privacy (SP)*, pp. 1450–1465. (https://doi.org/10.1109/SP40000.2020.00054)

Appendix

Focus Group Schedule

10 min	**1.**	**AI introduction** • Can you tell me what comes to mind when you hear the term "artificial intelligence"? • What do you know about how AI is used in the public sector in the Netherlands?
5 min	**2.**	**Example 01:** • The Dutch government developed a system to better detect welfare fraud. • The system created risk profiles of people supposedly more like to commit fraud by looking at things like their taxes, birthplace and education. • In 2019, the *toeslagenaffaire* showed that the system discriminated against certain groups. **Example 02:** • *Signalen Informatievoorziening Amsterdam* (SIA) is an online service where Amsterdam residents can submit complaints about public spaces and noise. • Residents can report any complaint on the same website page and AI categorizes it, sending it to the right department. • This service improves communication between the municipality and residents, and can help with the 250,000 misfiled reports per year in Amsterdam. Please keep these examples in mind for the coming questions.
20 min	**3.**	**AI & public sector** • What are important things to consider for the public sector when using AI? • Are there any specific concerns you have about the use of AI in the public sector? • How important are ethical aspects of AI?
10 min	**4.**	**Ranking** • What is the one most important thing for the public sector to prioritize when using AI? Write it down on a piece of paper. • Agree on a ranking with each other.
15 min	**5.**	**Tension examples** Not all priorities can be pursued to the same extent, because sometimes they conflict with each other. What is more important? Case 01 • AI can speed up the process of reviewing unemployment benefits, but needs access to a large amount of personal data to be able to work faster and accurately. • Should we allow the data to be used or not? Case 02 • Prisons use AI to select which individuals can be released early on good behaviour. Only the data of individuals with registered

		addresses are considered, but not others (e.g immigrants). Most of the time, the system's predictions are accurate. • Should we continue to use this system or not? <u>Case 03</u> • An AI system was used to identify potential residents interested in discussing sustainability with the municipality. You receive an invitation, but your friend David did not because he just moved to the city. • Should we send invitations to everyone, or only those who showed interest? <u>Case 04</u> • An AI-enabled virtual assistant is used to help answer citizens' general questions. It is no longer possible to call an information desk for help. The virtual assistant is available any time and responds quickly. • Should we only use virtual assistants or keep employees?
5 min	**6.**	**Question round** • Do you have anything to add?

Declaration of Authorship

I hereby declare that, to the best of my knowledge and belief, this Master Thesis titled "Ethical Implications of AI Use in the Public Sector: An Exploratory Research on Dutch Citizens' Perspectives" is my own work. I confirm that each significant contribution to and quotation in this thesis that originates from the work or works of others is indicated by proper use of citation and references.

Haarlem, 03 June 2023

Sarah Sherif Fathalla

Consent Form

for the use of plagiarism detection software to check my thesis

Name: Fathalla
Given Name: Sarah Sherif
Student number: 530667
Course of Study: Public Sector Innovation and eGovernance
Address: Schlossplatz 2, 48149 Münster
Title of the thesis: Ethical Implications of AI Use in the Public Sector: An Exploratory Research on Dutch Citizens' Perspectives

What is plagiarism? Plagiarism is defined as submitting someone else's work or ideas as your own without a complete indication of the source. It is hereby irrelevant whether the work of others is copied word by word without acknowledgment of the source, text structures (e.g. line of argumentation or outline) are borrowed or texts are translated from a foreign language.

Use of plagiarism detection software. The examination office uses plagiarism software to check each submitted bachelor and master thesis for plagiarism. For that purpose the thesis is electronically forwarded to a software service provider where the software checks for potential matches between the submitted work and work from other sources. For future comparisons with other theses, your thesis will be permanently stored in a database. Only the School of Business and Economics of the University of Münster is allowed to access your stored thesis. The student agrees that his or her thesis may be stored and reproduced only for the purpose of plagiarism assessment. The first examiner of the thesis will be advised on the outcome of the plagiarism assessment.

Sanctions. Each case of plagiarism constitutes an attempt to deceive in terms of the examination regulations and will lead to the thesis being graded as "failed". This will be communicated to the examination office where your case will be documented. In the event of a serious case of deception the examinee can be generally excluded from any further examination. This can lead to the exmatriculation of the student. Even after completion of the examination procedure and graduation from university, plagiarism can result in a withdrawal of the awarded academic degree.

I confirm that I have read and understood the information in this document. I agree to the outlined procedure for plagiarism assessment and potential sanctioning.

Haarlem, 03 June 2023

Sarah Sherif Fathalla