Deliberating AI

Why AI in the Public Sector Requires Citizen Participation

Master's thesis to obtain a MSc degree in Artificial Intelligence
by
Anna Dollbo
Student number: 1843672

July 11, 2023
Credits: 30 EC

Utrecht University
Faculty of Science
Department of Information and Computing Sciences
Princetonplein 5
3584 CC Utrecht

Project Supervisor (first examiner)
Dr. Dominik Klein

Second examiner
Dr. Johannes Korbmacher

Faculty of Humanities
Department of Philosophy and Religious Studies
Utrecht University
Janskerkhof 13
3512 BL Utrecht

Abstract

AI is a technology used across different societal domains due to its ability to increase operational efficiency by automating tasks and decision-making. AI is increasingly also applied in the public sector with the aim of making public administration more personalised, lean, and efficient. However, the technological advancement of AI also raises concerns regarding fairness, transparency, privacy, and human rights. Contrary to the perception of AI as a neutral tool, biases in its use and practices of surveillance suggest that it may have non-neutral impacts. Such impacts raise questions about AI's compatibility with democratic values. This thesis argues that if the use of AI in the public sector fails to uphold democratic values, its legitimacy is called into question based on democratic principles. The thesis investigates how AI in the public sector can impact the democratic values of equality, justice, and freedom. The investigation shows that AI should be understood as a political technology that risks disrespecting democratic values, raising questions about the legitimacy of its use in the public sector. The theory of deliberative democracy is employed to propose citizen participation as a way to address this. The thesis proposes the use of public deliberation as a means to determine how citizens would like to be governed by AI.

Keywords: AI, public sector, democracy, legitimacy, public deliberation

Contents

1. Introduction

Artificial Intelligence (AI) is a set of technologies increasingly used across diverse societal domains. AI's ability to process and analyse vast amounts of data enables the automation of tasks and decision-making, helping to improve the efficiency of organisational operations. In recent years, AI has increasingly been applied in the public sector with the aim of making public administration more personalised, lean, and efficient (Madan & Ashok, 2022). However, using AI also creates new risks and challenges regarding fairness, transparency, privacy, and human rights. Often conceived as a neutral tool, AI is perceived as able to achieve institutional goals and improve decision-making by objective reasoning (Green & Viljöen, 2020). However, the occurrence of algorithmic biases and pervasive surveillance seem to imply that AI might not have neutral impacts. This raises questions about AI's suggested value-neutrality and whether, in a democratic country, the use of AI in the public sector could impact democratic values like equality, justice, and freedom. If the use of AI in the public sector does not respect democratic values, their use might be illegitimate according to democratic legitimacy (Peter, 2017). This warrants an investigation into how democratic values might be impacted by the use of AI in the public sector and, if they are, how the use of AI can be deemed legitimate.

In this thesis, I will argue that AI is a political technology and that its use in the public sector necessitates citizen participation and legitimisation. I will use the theory of deliberative democracy to argue that the legitimacy of AI use in the public sector can be achieved through public deliberation. I will show that there are numerous decisions about where and how to apply AI in the public sector that can affect democratic values and hence citizens negatively. These decisions contain value-judgements about how society should be structured that need to be decided by the citizens through processes of deliberation and consensus-making. I will suggest what such a process could look like.

The key contributions of this thesis are twofold: firstly, to offer an analysis of the potential impact of AI in the public sector on democratic values, and secondly, to present a conceptual framework outlining the contours of a public deliberation process of AI in the public sector. The thesis is structured as follows: In Chapter 2, I will present the theoretical background. I will define AI and describe current and imagined AI use in the public sector and their perils and promises. I will also report on the current state of citizen participation in AI in the public sector. I will also introduce the theory and practice of deliberative democracy. Provided with this background, in Chapter 3, I start building my claim that AI in the public sector should be deliberated by arguing that AI is a political technology. I will do so by refuting the idea of technology as a neutral tool and showing that different levels of analysis can aid an understanding of AI's political nature. Having established AI as a political technology, in Chapter 4, I begin my analysis of the impact of AI in the public sector on democratic values by looking at the closely related values of equality and justice. In Chapter 5, I treat the impact of AI use in the public sector on the democratic value of freedom. In Chapter 6, I discuss theories of political legitimacy which provide different accounts of how AI in the public sector could be deemed legitimate. I also provide an analysis of different methods developed to increase accountability of AI use, which could potentially be used to discern whether

algorithmic decisions and AI solutions are legitimate. In Chapter 7, I discuss two common debates in AI and contemplate how they could influence public deliberation on AI use in the public sector. Finally, in Chapter 8, I give a practical account of how a public deliberation on AI in the public sector could be structured, provide a guideline for what should be deliberated, and discuss challenges to the process.

2. Background

2.1. Artificial Intelligence

Artificial Intelligence is a science and engineering practice in which different techniques are used to make a computer mimic human behaviour and cognition. AI systems consist of algorithms, a set of instructions programmed to solve a problem or perform a function (Merriam-Webster, n.d.). Algorithms use different methods for their operations, including mathematics, logic, statistics, and probability theory. Machine learning is a common subset of algorithms, which learn patterns and make predictions from data (Pumperla & Ferguson, 2019). Algorithms are human-designed, and given data and an objective by humans, algorithms carry out instructions to create a *model* that best fulfils the objective. The model can then be used to generate specific outputs such as content, predictions, recommendations, or decisions that influence the application environment of the algorithm (*Artificial Intelligence Act*, 2021/206). An example is fraud detection in government transactions which uses historical transaction data to create a model of fraudulent patterns that can be used to analyse real-time transactions and flag suspicious activities for further investigation.

AI aims to replicate cognitive abilities, including reasoning, knowledge representation, planning, learning, natural language processing, and perception (Russel & Norvig, 2021).
These are approximated in computing technologies like computer vision (the ability to derive meaningful information from images, videos, and visual input), natural language processing (the ability to process and analyse natural language), speech recognition (the ability to identify words spoken aloud and convert it to text), knowledge representation (the ability to organise and structure knowledge in a meaningful way) and pattern classification (the ability to recognise and categorise data patterns based on previous examples) (Medaglia & Tangi, 2022).

However, AI is not just a set of technical methods. When considering the actuality of AI, one cannot separate the computations by the algorithms from the social context in which it is applied (McQuillan, 2022). Crawford (2020) writes that AI is always connected to broader structures and social systems, making it both a technical and social practice related to institutions, infrastructure, politics and culture. AI reflects and produces knowledge of the world and social relations.

2.2. AI in the public sector

2.2.1. Current uses

Within the public sector, governments act as regulators of AI technologies to ensure their safe development. A less discussed dynamic is that governments are also users of the technology. Governments worldwide are already adopting AI technologies at local, regional, and national levels. Currently, the technologies are deployed in various policy areas such as public order and safety, defence, education, environmental protection, housing and community amenities, economic affairs, health, and social protection (Sousa et al., 2019). Some examples of uses of AI technologies in the public sector are predictive policing tools used in Germany, The Netherlands and United

Kingdom, automated immigration processes in Canada, optimisation of employment services in Poland, virtual assistant Alex to help with tax services in Australia, and AI technologies to analyse aerial imagery to spot undeclared properties to prevent tax fraud in France (Kuziemski & Misuraca, 2020; Neslen, 2021; Pannett, 2022; Sousa et al., 2019).

According to the report *Global Trends in Government Innovation 2023* (2023) by the OECD, these technologies have the potential to make the public sector smarter, which is here interpreted as agile, efficient, and user-friendly. According to the report, these features of AI technology are also said to increase the trustworthiness of the public sector. Opportunities for AI in the public sector are mainly seen as belonging to three areas: (1) enhancing the internal efficiency of public administration, (2) improving public administration decision-making, and (3) enhancing interaction between citizens and government by better and more inclusive services and the improvement of citizen participation in the public sector's activities (Medaglia et al., 2021). The adoption of AI in governments is seen as a way to modernise the public sector by using different opportunities of the technology to make public institutions more effective and adaptive to change (Araya, 2015).

Many governments have developed strategies for how to use AI in their work. Currently, 69 countries have AI strategies and policies in place (*OECD's Live Repository of AI Strategies & Policies - OECD.AI*, n.d.). Ossewaarde & Gulenc (2020) analysed AI strategy papers for the United Kingdom, Germany, and The Netherlands to uncover political narratives implied in the AI technologies employed and what type of national ambitions they reveal. Their analysis showed that all three countries mythologise AI as a benevolent force of national progress that can help realise long-standing political ambitions, especially in the context of global competitiveness among nations. They noticed that none of the AI strategy papers included a vision of democracy or a strengthening of democracy. They also argue that all strategies hide the fact that AI technologies not only solve problems but also create new ones.

The positive picture painted by national AI strategy papers is counteracted by researchers who have identified and warned about the many risks and challenges of AI adoption in the public sector. These include the risk of widening of social inequalities, infringement on citizens' privacy, problems of transparency and, by extension, accountability, exclusion of certain actors, increased complexity of analysis, new regulatory requirements, dehumanisation of daily activities, job displacement, technology dependence and obedience, loss of control, and AI paternalism and dominion (Medaglia et al., 2021; Valle-Cruz et al., 2019; Wirtz & Müller, 2018).

2.2.2. Empirical results and participation

Despite the many opportunities and risks identified, empirical research on the actual impacts of AI technologies in the public sector is still lacking. Medaglia & Tangi (2022) performed a survey investigating the perceptions of drivers, features, and impacts of AI applications in the public sector by public managers in the EU member states. Their results showed that public managers had an overall positive mean evaluation of the perceived impacts of AI technologies. This perceived positive impact is mainly on internal operations, while public managers give lower scores to more long-term impacts on social value and well-being, openness and inclusiveness. However, their

results also showed a lack of transparency on all aspects of AI applications. They also found a significant lack of citizen involvement in all stages of the development process (planning, piloting, and deployment). These are relevant findings considering a recent poll on public opinion on AI use by governments which indicated that citizens in polled countries are seriously concerned about the application of AI for national security (*New Poll: Public Fears Over Government Use of Artificial Intelligence*, n.d.). Even though increasing citizen participation is said to be one of the main opportunities for AI in the public sector, it is still lacking in practice.

As states have duties and obligations under international law to respect, protect and fulfil human rights, the public sector has a higher duty of care in using and developing AI technologies than the private sector. A recent recommendation report published by the Council of Europe Commissioner for Human Rights emphasised how the private sector's narrative of AI as so highly technical and inscrutable that it cannot be effectively controlled and regulated disincentivises senior policy levels to engage comprehensively with the technology and its potential risks to human rights (*Human Rights by Design*, 2023). The report urges member states to hold regular public consultations about AI, involving not only experts, industry, researchers, and academia but also the wider public and representatives of the most affected groups. It highlights the importance of member states developing AI literacy and awareness to ensure that both those governing AI systems and those governed by them understand the technologies and their impacts on human rights.

2.3. Deliberative democracy

2.3.1. Citizen influence and political legitimacy

Deliberative democracy is one of the most prominent theories of democracy today (Rostboll, 2008). It is the ideal that citizens should participate more in decision-making procedures by coming together and discussing the political issues they face on the basis of equal status and mutual respect. Subsequently, deliberative democracy also holds that the decisions reached through this public deliberation should have priority over economic and social powers. Deliberative democracy is a more demanding ideal than the common model of aggregative democracy, which is decision-making through the aggregation of individuals' preferences, hence voting (Peter, 2009). Deliberative democracy seeks to increase the citizens' influence beyond voting in elections to impact instead the political choices made by the government through deliberative participation in the decision-making process. Central to the idea of deliberative democracy is that political decisions should be based on good *reasons* and that the subjects of those decisions, the citizens, are within their rights to know what those reasons are and to "have their say". By being able to have a say in the policies that are created, deliberative democracy seeks to increase citizen freedom and the legitimacy of government since, if created deliberatively, the only laws we as a people have to abide by are those we have given to ourselves (Neblo, 2015).

2.3.2. Values of Deliberative Democracy

The idea of mutual respect is central to deliberative democracy. In practice, this means that the participants in deliberation are expected to actively listen to and try to understand the meaning of a

speaker's statements, see their motives and consider their argument as opposed to viewing them as objects to be dismissed (Bächtiger et al., 2018). Inclusion is another ideal in deliberative democracy, commonly meaning that all those whose interests are at stake in the decision should have a say in the discussion. Another important ideal is equality of communicative freedom, meaning everyone should be equally free to express themselves and their points of view. This kind of mutual communication involves a weighting and reflection on preferences, values and interests for issues of common concern under conditions of mutual respect and equality.

Deliberation is also valued for the byproduct it creates, where deliberation enables people to voice their grievances which in turn fosters mutual understanding and community-building (Estlund & Landemore, 2018). For many deliberative theorists, these outcomes are more important than reaching a consensus from a deliberation process. Gutmann et al. (2018) point out that consensus is not always desirable. Suppose consensus is proclaimed when there are deep and persistent disagreements. In that case, it could lead to the marginalisation of dissenting voices and frustrate efforts for future deliberation that result in more just outcomes. They argue that deliberation should instead be seen as a method for reaching mutually respectful agreements, even if they fall short of consensus or only satisfy one particular conception of justice, for example. This is a more realistic approach as democratic politics is about "publicly defensible compromise among a diverse group of free and equal people who routinely disagree about what laws and public policies are morally best". Even so, reaching a consensus about a problem is still an important goal for deliberation since this lends legitimacy to the decision and further incentivises public officials to execute it (Hartz-Karp et al., 2018).

2.3.3. The epistemic value of deliberation

Deliberation has epistemic value for getting to the correct or right choice or as close to it as possible (Estlund & Landemore, 2018). A correct decision can mean many things: the objective truth about a matter, which can be both facts or morals or an intersubjective, culturally dependent and temporary construct. From an epistemic point of view, the reason for deliberation is to figure something out, whether that be the truth, a correct decision, or a socially useful answer. Landemore (2013) argues that the logic of why deliberation would produce good knowledge or correct decisions can be found in the Diversity Trumps Ability Theorem by Lu Hong and Scott Page. They propose that in some circumstances, a randomly selected group of cognitively diverse individuals can better identify the best outcome than a collection of experts with the same number of members (Hong & Page, 2004). This implies that cognitive diversity - how people think about a problem - results in collective wisdom. Critics of this theory argue that it does not hold for all kinds of deliberation landscapes and that problems might arise when transferring the results from the model to a real-world setting (Grim et al., 2019). Results from agent-based modelling approaches indicate that diversity is most beneficial in larger deliberation groups and that a mixed group of experts and non-experts is better than a homogenous group.

2.3.4. Deliberative democracy in practice

Deliberative democracy's ideals are aspirational: they cannot all be fully achieved in practice but provide a good standard (Bächtiger et al., 2018). The practice of deliberative democracy is public

deliberation, which can be described as a public discussion that aims to create collective solutions to difficult social problems (Blacksher et al., 2012). This practice can take various forms, such as citizen juries, national issue forums, deliberative opinion polls and participatory budgeting. A literature review by Carpini et al. (2004) of empirical research on the individual and collective benefits of public deliberation has given insights into the usefulness of the practice. The results show that contrary to claims that citizens lack interest in discussing public issues, the research shows that enough citizens engage in public discussions to understand the role of deliberation in democratic politics. Another result is substantial evidence from social psychology that deliberation can lead to individual and collective benefits. These benefits include empathy for the other, a broader sense of one's interests, a more enlightened citizenry concerning their own and other's needs and experiences, a more politically engaged citizenry, and an increased competence to resolve deep conflict.

3. The political nature of AI

3.1. AI as a neutral tool

3.1.1. Instrumentalism

One of the most common understandings of technology portrays it as an instrument for human ends. This characterisation of technology is called instrumentalism in the philosophy of technology, which investigates the nature of technical artefacts, technological knowledge, the philosophy behind the design and engineering of technologies, and norms and values in technology (de Vries, 2018). Following the instrumentalist approach, AI is often also posited as a neutral tool made by neutral actors (Green & Viljöen, 2020). This means that the technology itself is not seen as value-laden. Rather it derives all its impacts from its uses by humans. The instrumentalist approach holds that, except for in human uses, technology itself cannot change or impact society (Verbeek, 2022). Pitt (2014) has captured this idea in the Value-Neutrality Thesis (VNT), which states:

Technological artefacts do not have, have embedded in them, or contain values.

This value-neutral conception of technology means that instrumentalism mostly focuses on descriptive practices of technical analyses. The main focus of discussion for technology then becomes what ends it should serve and what scientific principles can be applied to obtain the ends. AI in the public sector is posited as a tool to make the government more efficient, agile, and user-friendly. In the instrumentalist view, the utility of technology to solve problems becomes the only normative criterion for evaluation, placing "trade-offs" at the centre of the discussion (Feeberg, 1991; Swer & du Toit, 2020).

3.1.2. The value-free ideal

The proposed value-neutrality of technology in instrumentalism is not just seen as a descriptive fact about technology but also an ideal to be realised through science (Swer, 2014). Green & Viljöen (2020) describes how algorithms often are perceived as tools that can make "objective" and "neutral" decisions. Instrumentalism hence aligns with what is called the "value-free ideal" in science, which states that social, ethical, and political values should not influence the reasoning of scientists and that "good" science relies on suppression of self, which is believed to lead to an objective outcome (Douglas, 2009). In the instrumentalist view, technology is seen as the result of science. The value-free ideal holds that scientific knowledge deals with facts and not values. Technology, being applied science, is seen as applied facts and is therefore considered to be outside the realm of values (Swer, 2014).

This demarcation between facts and values helps instrumentalists do technological analysis by separating what is considered relevant (objective issues and data) from what is considered irrelevant (the subjective and affective) (Swer, 2014). This means that any ethical or social issues arising from the technological operations are seen as being outside of the technology, belonging to the realm of values, not the technology itself. Hence, any investigations into other aspects of how the

technology might contribute to certain value-based phenomena are deemed superfluous. Pitt (2000) argues that technology should not be analysed through ethical and political perspectives since “tools and technical systems are inherently ideologically neutral” (p.72).

3.2. Political nature of AI

3.2.1. Criticism of instrumentalism and the value-free ideal

In insisting that it only deals with facts, instrumentalism tries to remain descriptive, not normative. Douglas (2009) criticises this value-free ideal by arguing that scientific research is inherently value-laden and that values have an important role to play at various points throughout the research process. Values are involved in the selection of the research question, in the interpretation of data, and in how the scientific findings are ultimately applied. Douglas argues that the value-free ideal misinterprets many values for facts and claims that acknowledging and engaging with the values is the best way to increase scientific research's objectivity and reliability. Furthermore, she warns that the value-free ideal can lead to a lack of accountability. This is also true for AI development: claiming the models created by the algorithms remain value-neutral - despite all the value-laden choices made in the design process - can lead to algorithmic harms that could have been anticipated if the technological analysis had been expanded.

By positing technology as a mere tool, the focus of analysis becomes the micro-level. For AI, this means an analysis of the algorithm per se, for example, the code or the model. This might be a useful level of analysis to reason about certain mathematical or epistemological aspects. However, it does not say much about the technology beyond this piece of code. Another level of analysis looks at technology at the system level or collectively as a group of technologies. This level includes the people behind the algorithm, the data used by the algorithm, the model that the algorithm creates, and the countless interactions of people and systems with the technology over time in its understanding of technology. In this macro-perspective, it becomes difficult to talk about an algorithm as neutral except for its use by people. How an algorithm impacts a person's life depends on technological choices of the algorithm's design. The value-free ideal in science that creates algorithms in this neutral view requires each technological artefact to be historically and socially isolated from its development process. This leads to a lack of emphasis on the social consequences of technical choices when doing technological innovation (Swer & du Toit, 2020).

3.2.2. A complementary approach - technological materialism

In contrast to the instrumentalist approach, Winner (1977) argues that technological development often aligns with current prominent moral and political systems. He argues that positing technology as neutral allows new technologies to unreflectively merge with pre-existing techno-structures, which limits the types of society that can be brought about. When instrumentalists consider technologies as facts, the nature of technological development becomes a technical issue, preventing technology from being seen as a political or social matter. When technology is only considered a technical issue, it becomes difficult for non-technicians to judge it, and any decisions about the technology and its development is hence seen as best left to experts.

Instrumentalism tends to show a lack of critical assessment towards technological experts' influential and unregulated roles in societal decision-making (Shrader-Frechette, 1994).

According to Winner, the instrumentalist understanding of values as external to technology obscures three important facts. First, the instrumental concepts of use and user are pointless when referring to technological systems rather than tools. In a scenario where an algorithm provides certain benefits, it would be difficult to discern who the user would be and whether their intentions matter in the outcome. Second, the structure for technological systems might not be chosen so much as necessitated by the physical requirements of their operations. An example is how image processing requires a certain algorithmic architecture that can process images. This architecture receives and interprets information in a particular way, which can introduce biases or raise privacy and security issues. Third, the instantiation of technological systems requires a material restructuring of society. An example is how AI necessitates data collection infrastructures for the environments in which AI should work. This means that the consequences of technological development are not primarily conceptual but also material. They require humans to adapt their behaviours to the systems, making some behaviours more permitted than others. Hence, this view can be called technological materialism (Swer, 2014).

In technological materialism, technological systems can bring about transformations that concern all aspects of the social sphere; social relations, political systems, moral norms, and cultural forms (Swer, 2014). In this sense, technology both requires legislation and also becomes a form of legislation on how humans should operate and exist in the polis. This means that technology is a political phenomenon (Winner, 1977). If decisions of technological development to some extent influence the social sphere, then those decisions and transformations are matters for public debate and political action. In this view, the use of AI in the public sector is not a private technical matter to be left to experts but a public matter that is up for deliberation.

3.3. AI as a topic for deliberation

Technological development does not only pertain to questions of “what?” but also of “why”? In the instrumentalist view, the “what” becomes the descriptive “facts” about the technology and the “why” becomes the specific problem it was optimised to solve. In the technological materialist conception of technology, the “what” pertains to the larger system and its implications on the individual and societal level, and the “why” is a question those who develop technology and the society at large can answer. As a democratic government has a role to fill both for its citizens and the democratic institution itself, it becomes important to investigate how the “what” of AI in the public sector impacts the democratic values and principles it is bound to uphold. This requires both a technical instrumentalist understanding of how the algorithms work and a technological materialist understanding of how AI impacts social structures. Only when such a broader analysis is done can we accurately judge whether the “why” of AI in the public sector is legitimate.

From a deliberative democratic perspective, decisions made in the public sector should be based on good reasons that the public can agree with. The most legitimate decisions are those that the public has contributed to through deliberation. This means that the

“what” of AI in the public sector can be deliberated to reach a legitimate “why”. By looking into how AI in the public sector can impact the democratic values of equality, justice, freedom, as well as democratic legitimacy, it becomes possible to argue that the potential impact on both the democratic institution and the citizens is serious enough that the public should have a say in the matter. In the next three chapters, I shall carry out this investigation and show what could be deliberated about in a public deliberation on the government's AI development and use.

4. Equality and justice

4.1. Definition of equality and justice

4.1.1. Equality

Equality has, since the French Revolution, been one of the leading ideals in politics but also one of the most contested (Gosepath, 2021). Equality signifies a comparative relationship between people with the same qualities in at least one respect. Some people understand equality as sameness with respect to the possession of certain goods, such as resources, welfare, or capabilities for example. Another view of equality sees equality as relational, where the goods are social relations of symmetrical or reciprocal authority, recognition and standing (Anderson, 2012).

Considerations of equality often begin with an analysis of social hierarchies. Social hierarchy means persistent group inequalities sustained by laws, norms, or habits (Anderson, 2012). These hierarchies are reproduced over time and create classes of people that relate to each other as superiors and inferiors. The social hierarchies are based on attributes of group identities such as race, class, ethnicity, gender, religion, sexuality, age, citizenship status or language. Egalitarians generally reject social hierarchies based on three perspectives related to the three domains of values: the good, the right, and the virtuous (Dewey, 1981). Firstly, they argue that social inequalities are bad for people, both those in inferior positions and those in superior positions and society as a whole (Anderson, 2012). Secondly, they argue that inequalities are morally wrong since it is unjust to those placed in inferior ranks. Thirdly, they argue that social inequality is vicious since it corrupts the characters of superior and inferior alike. Out of three, the dominating judgment for why inequality is wrong is based on the claim of justice.

4.1.2. Justice

Justice can be seen as a moral concept used in relation to another person and which concerns "what we owe to each other" (Miller, D., 2021). In practice, justice often refers to fairness in the distribution of benefits and burdens to persons in society, hence distributive justice (Anderson, 2012). This means that conceptions of justice need to specify several ideas: what is regarded as a "fair" distribution, what are the benefits and burdens to be distributed, what are the necessary conditions for being a person for whom justice matters, and what temporal and spatial scope justice has. Justice helps establish what inequalities should be corrected, what inequalities are justified, and on what grounds. Examples of principles of distributive justice are strict egalitarianism, John Rawl's account of fair equality of opportunity, and luck egalitarianism.

Strict egalitarianism is the distributive principle that everyone should have an equal amount of a certain good or resource to minimise overall inequality across individuals (Arneson, 2013). Strict egalitarianism is commonly rejected on multiple grounds. One ground for rejection is that strict egalitarianism lessens incentives for wealth-accumulating activity since everybody gets the same regardless of merit or effort. Another rejection is that strict egalitarianism does not factor in responsibility in its distributive principle. This would mean that resources would be continuously

transferred to individuals who squander their resources, which for some is deemed unfair to those who do not.

John Rawls answers the criticism of strict egalitarianism by allowing for inequalities but only on certain grounds. His approach still incentivises people to work and obtain more resources than others, but only if it satisfies the difference principle: any inequalities must improve the situation of the most disadvantaged individuals (Rawls, 2009). He also holds that positions that allow one to obtain more resources must be open to all under fair equality of opportunity. This means that everyone with the same talent and eagerness to use that talent must have equal opportunity to do so. Moreover, he requires that all individuals be able to develop their talents to the fullest extent. This would mean a redistribution of resources, so everybody has an equal starting point, showing Rawl's account's relation to strict egalitarianism.

Luck egalitarianism responds to the lack of responsibility in strict egalitarianism by aiming to distribute resources and opportunities based on a conception of luck. According to luck egalitarianism, only those inequalities that come from factors such as natural talents or social circumstances - matters of "bad luck" are considered for redistribution (Dworkin, 2000). However, any inequalities resulting from a person's choices - so-called "option luck"- are considered acceptable. This could pertain to an individual's career choices, educational decisions, or lifestyle preferences.

4.1.3. Investigating AI through equality and justice

Given these different conceptions of equality and justice, an investigation into how AI in the public sector impacts these ideals would need to consider how AI treats people differently given different conceptions of "what we owe one another". In a democracy, there is the idea that people should have the opportunity to advance their interests equally. This ideal of public equality ensures that people can conceive that they are being treated as equals (Christiano & Sameer, 2022). This ideal also posits limits to democratic decision-making, as public equality requires that liberal and civil rights be protected. Suppose the benefits and burdens of algorithmic outcomes are not distributed equally. In that case, it is important to investigate whether this distribution can be considered fair and what AI can and cannot do regarding inequality. When choosing an algorithmic solution for a problem, there are many ways that bias, and therefore values, can enter the algorithms and potentially cause discrimination and harm. In this chapter, I investigate how algorithms in the public sector could mean violations of public equality and fairness.

4.2. Algorithmic bias

4.2.1. Understanding bias

Bias for algorithms can have several meanings, but algorithmic bias generally is any systematic deviation in output, performance, or impact relative to some norm or standard (Danks & London, 2017). The effects of these deviations differ, making it possible to speak of the algorithms as morally, statistically, or socially biased, depending on the normative standard the algorithm was

based on. Statistical bias is when an estimate deviates from the true value (Piedmont, 2014). Statistical bias can occur in algorithms in different ways. For example, if the model created of the data an algorithm was trained on deviated from the true distribution of the data. When used for prediction, this can create different problems. An example is how during the Covid-19 pandemic, many predictive algorithms used in banking had an increased inaccuracy as a result of people's changed banking habits during this time (Anderson et al., 2021). Since the data changed, the model no longer accurately represented the true values.

An algorithm can be morally or socially biased if it depends illegitimately on specific attributes deemed irrelevant to the problem, such as gender or social group membership. An example is an algorithm that presented different job opportunities for men and women, with women being shown fewer ads for STEM positions than men (Lambrecht & Tucker, 2019). It is important to note that not all statistically biased algorithms will be morally or socially biased. This algorithm might not necessarily be statistically biased as women only make up 28% of the STEM workforce (Piloto, 2023). However, according to a normative standard that gender should not influence the opportunities for work, this algorithm can be considered biased. This shows that algorithms that are not statistically biased might still be morally or socially biased if they depend on illegitimate attributes according to normative criteria (Fazelpour & Danks, 2021).

4.2.2. Sources of bias

Developing new algorithms involves several points of unknowability in the design process, which requires the designer to make choices. These are the places where potential biases can get embedded into the algorithm (Stelmaszak, 2021). The first point of entry for biases in algorithms is in the problem formulation. This requires thinking about the goal that the algorithm should be used to achieve (Fazelpour & Danks, 2021). Many popular decision-making algorithms are learning algorithms that create a statistical model based on historical data to predict the output of new, unforeseen data. In order to create these predictions, they need to be given a target, a goal which they should maximise or optimise for. The choice of the target variable can create bias in the system if it does not accurately reflect or present a valid solution to the real-world goal (Mitchell et al., 2021). An example is how the goal of higher education, in general, might be reduced to simply maximising the future grade point averages of admitted students (Kleinberg et al., 2018). This example highlights that often, the things that matter cannot be accurately measured, meaning that algorithms must rely on proxies. These proxies are sometimes in themselves sources of bias. An example of that is an algorithm used for predicting which patients would benefit from high-risk care management by hospitals. The developers used insurance claims as a proxy for health risk. This led to discrimination against Black patients who were not offered help until they were much sicker than White people. This result came from the fact that Black patients often get much sicker before making an insurance claim than White people, making insurance claims a poor proxy for health risk (Obermeyer et al., 2019).

Data is another source of bias. Data from the real world reflects the biases already existing in the real-world system, which then becomes captured in the statistical model created by the algorithm. This is captured in the slogan “bias in, bias out”, considered common folk wisdom in the machine learning community (Rambachan & Roth, 2019). Bias in data can be due to data labelling

processes, where data can be labelled according to subjective opinions with normative implications. An example is how the image dataset ImageNet's Person category has subcategories with labels such as "Bad Person, Call Girl, Drug Addict, Convict, Crazy, Failure, Fucker, Hypocrite, Jezebel" ascribed to images of people (Crawford & Paglen, 2021). In other cases, the absence of perspective and context in data can be problematic. An example is a photo of an Israeli soldier holding down a Palestinian boy while the boy's family tried to remove the soldier, which was labelled "People sitting on top of a bench together" (Katz, 2020). Bias in the data can also happen because of limitations and bias in the data collection methods. The time of the data collection or the specific sample taken could lead to representation bias where the data is not representative of the relevant population (van Giffen, 2022). Such representation bias can happen if, for example, only data from a certain demographic or cultural group is collected, which can cause the algorithm to underperform for the under-sampled groups. An example is how facial recognition algorithms have been proven to perform worse on black people in general and black women in particular because of undersampling in the training data (Buolamwini, 2017).

Bias in the development process can also happen in the modelling and validation phase. This is often an iterative optimisation phase that tries to ensure the particular model "fits" the data relative to some success criteria (Fazelpour & Danks, 2021). The metric denoting success will favour one performance over another and so is not value-neutral. This often involves making choices about tradeoffs, which can have ethical consequences when there are independent, irreducible objectives that need to be satisfied simultaneously (McQuillan, 2020). Bias can also result from misunderstandings of the algorithm outputs when applying predictive algorithms to decision-making contexts. The algorithms' predictions are often purely observational: they might describe the problem's relevant features but do not necessarily explain the root cause. If the algorithms' predictions are used as credible explanations of the problem, it could prevent relevant actors from addressing the actual root cause.

4.3. Algorithmic harms

Bias in algorithms can harm individuals and groups in society. Why this is wrong is related to which normative standard one applies. Differences in the treatment of groups and individuals in society relate to questions about the distribution of benefits and harms, and therefore to questions of justice. Justice primarily concerns the treatment of individuals, but as seen in the examples above, algorithmic harms often affect certain groups of individuals. Unfair treatment by an algorithm then would mean that an algorithm treats an individual differently as a cause of them having a specific trait (Hedden, 2021). This unfair treatment could lead to harm or negative consequences for the individual.

Barocas et al. (2017) distinguish between two different types of algorithmic harms: allocative and representational. Allocative harms are when opportunities or resources are withheld from certain people or groups. This type of harm is often immediate, easily quantifiable, discrete, and transactional. An example of allocative harm is an algorithm used to allocate educational resources that disproportionately favours schools in wealthier neighbourhoods, leading to fewer resources for schools in underprivileged neighbourhoods. Representational harms are when certain people or

groups are stereotyped and depicted in a discriminatory way. These harms are long-term, difficult to formalise, diffuse, and cultural. An example of representational harm could be a natural language processing algorithm that exhibits bias in its language generation, producing sexist or racist outputs that reinforce harmful stereotypes. Allocative and representational harm can also inform one another so that allocations of resources lead to a perpetuation of inequalities which might enhance stereotypes that some people are more deserving than others. Likewise, inaccurate representations of groups could lead to biases that inform decision-making processes. This relates to the association between recognition and distribution, which will be touched upon in the section on freedom. In the coming parts, the focus will mainly be on allocative harms.

4.4. Mitigating bias in algorithms

4.4.1. Algorithmic fairness

That algorithms can be considered unfair means that there is a gap between the beneficial results promised by AI applications and the actual impact and consequences of the systems as they are deployed in society (Dobbe et al., 2021). This divide can be described as the sociotechnical gap, which is "the great divide between what we know we must support socially and what we can support technically" (Ackerman, 2000, p.180). One suggestion on how to fill the sociotechnical gap and thus increase fairness and equality in (or through) algorithms is to apply different mathematical criteria to analyse and manipulate algorithmic outcomes (Dobbe et al., 2021). This approach is called algorithmic fairness and involves developing mathematical definitions of fairness, auditing the algorithms for violations of those definitions, and mitigating unfairness by applying certain policies (Green, 2021).

There have been many different fairness definitions or metrics developed. These different metrics can be distinguished in two ways: they are either observational or causality-based criteria, and they are either criteria based on group membership or individuals (Castelnovo et al., 2022). The most commonly discussed fairness criteria are the statistical criteria of fairness for groups. These require that certain relations between predictions and actuality are the same for each group (Hedden, 2021). The groups usually considered are those pertaining to characteristics that are protected or sensitive attributes from non-discriminatory law, such as race and ethnicity, gender, religion, age, disability, and sexual orientation (Lee et al., 2020). Some of the most common statistical group fairness metrics are demographic parity, equalised odds, and calibration within groups.

4.4.2. Example setup and notation

To discuss how different fairness metrics relate to distributive principles, it is first necessary to distinguish between prediction and decision. For an allocative algorithm used in the public sector, the decision task concerns the distribution of a certain resource. For example, an algorithm might be used to predict whether an unemployed person will still be unemployed after six months based on their work history, previous periods of unemployment, job market conditions, education level and other factors. They might distribute different resources to the claimants based on their expected employment status after six months, such as different job programs. The decision task is

the distribution rule which states: allocate an amount R of the resource (the job programs) to the claimant X (the unemployed) if X has attribute Y (unemployed after six months). In the example of the unemployment programs, it is unknown how long the person will be unemployed, which necessitates a prediction of Y. This is the prediction task. The attribute Y (employment status after 6 months) is predicted from X's (the unemployeds) observed attributes V. The observed attributes V are further separated into attributes A that are protected (eg. gender, race, age) and unprotected attributes B (education level, previous work history etc.). An algorithm will hence give a prediction $\widehat{Y}$ that hopefully corresponds to the true value of Y. If the prediction $\widehat{Y}$ is plugged into the distribution rule so that $\widehat{Y}$ is taken to be Y, this constitutes prediction-based decision-making. In other times, the predictions and the decisions might be separated so that, for example, a public administrator is given the prediction $\widehat{Y}$ and, based on that, decides what Y should be in the distribution rule. What is distributed is hence different for the prediction task and the decision task. Whereas in the decision task, what is distributed is the actual resource (the job programs), in the prediction task, what is distributed is prediction errors (which is when $\widehat{Y}$ does not correspond to Y) (Kuppler et al., 2021).

4.4.3. Equalised odds and equality of opportunity

If the job program allocation algorithm were analysed with the equalised odds fairness metric, it would be deemed fair if people from different groups of the protected attribute A have the same true positive rate and true negative rate, hence the same prediction error rate. True positive rate means that all the people who were employed after six months (Y) had the same probability of being predicted to be employed after six months ($\widehat{Y}$), irrespective of group membership (A). True negative rate means that all the people who were not employed after 6 months (Y) had the same probability of being predicted not to be employed after 6 months ($\widehat{Y}$), irrespective of their group membership (A). This means that people with the same underlying qualification (Y), as determined by their attributes (V), should have equal opportunities to receive the resource R (equal chance of $\widehat{Y}$ being Y). Hence, equalised odds as a metric roughly maps onto the distributive principle of Rawl's fair equality of opportunity. This criterion would be violated if, for two values of the sensitive attribute A : a and a *, if people from group a were more often predicted not to be employed than people from group a * despite both of the groups in this case actually being employed after 6 months (Y). However, equalised odds allows for different treatment of people with different qualifications (Y) (Kuppler et al., 2021). This means that it allows for different prediction error rates between people who were unemployed or employed after six months. The algorithm might hence work worse for people who have a lesser value of Y, which seems to mean that the algorithm does not work as well for different people, which violates the ideal of public equality. If the prediction task informed the decision task, then this would mean that despite people having the same qualifications, the different error rates in the algorithm's prediction could still lead to resources being given inconsistent with the distribution rule of the resource. Hence, equality of opportunity can only be partially realised through the equalised odds metrics.

4.4.4. Demographic parity and luck egalitarianism

Another metric is demographic parity, which in contrast to equalised odds, does not care about the true value of Y (whether people were employed or not after six months), but rather is a metric for the decision task which informs which prediction model is chosen (Kuppler et al., 2021). Demographic parity prefers prediction models where the predicted outcome $\widehat{Y}$ is the same for all groups of the protected attribute A. This means that all individuals, regardless of group membership, have the same probability of receiving the resource R (job program). At first glance, this criteria looks similar to strict egalitarianism since it distributes the resource the same across all groups. However, strict egalitarianism does not require an analysis of group membership. It distributes a resource the same to all individuals. Demographic parity requires that the acquirement of the resource R should not depend on outcomes Y that are due to the sensitive attribute A. If a person is unemployed because of their group membership A, which they did not choose, and despite having the attributes B (education, good resume) that should lead to a positive outcome of Y, then this can be seen as a matter of bad luck. Hence, demographic parity can be mapped loosely onto luck egalitarianism. By applying demographic parity to the algorithm, it becomes possible to mitigate the societal bias of unequal opportunities that are due to a person's group membership. Demographic parity is hence a fairness metric that seeks to alter the status quo, as most algorithms are not fair according to this principle (Castelnovo et al., 2022).

4.4.5. Calibration and the role of algorithms in fairness

Some would argue that it is not the prediction algorithms' job to alter the status quo. This would force the prediction algorithm to perform the decision task. Kuppert et al. (2021) argue that the prediction task should only be concerned with painting a picture of how the world actually is based on the available data. It is then the decision task's mission to use this picture of how the world is to decide how the world should be. Hedden (2021) advances a similar argument and claims that calibration within groups is the only fairness metric necessary for fairness. Calibration within groups refers to the alignment between predicted probabilities and observed outcomes for specific subgroups to ensure that the predicted probabilities accurately reflect the likelihood of positive outcomes within each group (Kleinberg et al., 2016). This means that if in group A, the algorithm predicts that 70% will be employed within six months (positive prediction $\widehat{Y}$), then the ratio of people from group A who are actually employed within six months (true value of Y) is also 70%. By devising a thought experiment, Hedden (2021) shows that calibration within groups is the only criterion not violated by a manifestly fair and unbiased algorithm. This is true even when there are equal base rates (equal distribution of Y for all groups). Based on this, he argues that it is not necessarily the algorithm that is unfair when used to make distributional consequences, but rather the decision task or the background conditions of society.

By arguing that it is the use of the predictions that are unfair and not the predictions themselves, Hedden ascribes to an instrumentalist understanding of technology by looking at the algorithms as the sole artefact to evaluate the fairness of. The perfect thought experiment set up by Hedden is, of course, unrealistic, which he concedes, and he is not arguing that it is the only metric that can be used for fairness but that it is the only necessary one. The primary goal of calibration within groups

is to assess the accuracy and reliability of predictions within groups. It can hence be argued to be more of a performance metric than a fairness metric. This is also in line with the instrumentalist tendency to want to distinguish between facts and values, where fairness is taken to be how well the algorithm represents reality for different groups. Instrumentalists, like Hedden, are concerned with algorithms' epistemological status and prefer to evaluate them on the knowledge they create.

However, even if calibration within groups is the only criterion necessary for fairness in an instrumentalist view, in a technological materialist view, the whole system is taken into account. It is common for predictions to feed directly into the decision tasks, which raises the question of whether the algorithm should predict outcomes that align with the society we want to create rather than create an accurate picture of a highly unequal reality. One could argue that, in contrast to private sector data scientists, the public sector has the ability to change society in a specific direction. Algorithmic fairness measures beyond a well-calibrated algorithm could potentially be a means to that end.

4.4.6. Fairness metrics and situational contexts

If algorithmic outcomes should help change society in a way deemed more fair, the question then becomes how to decide what that reality should look like, and based on that what fairness metric to evaluate the algorithms from. As discussed above, different metrics loosely map to different distributive justice principles, which might aid a discussion on what fairness metrics people deem appropriate. More than just deciding on one specific fairness metric, it would be important to consider the application context. Equality of odds might be a better metric when selecting candidates for a job interview, for example, but in matters of civil justice, one might want to use demographic parity, such as for airport security checks, so that no group is over-examined because of an algorithm (Binns, 2017). Different characteristics of the decision context influence the selection of a morally suitable fairness metric (Loi & Heitz, 2022). Different criteria have different advantages, and deciding the desirable outcome requires inevitable trade-offs between objectives of interests as it is impossible to simultaneously satisfy all mathematical definitions of fairness (Kearns & Roth, 2019). Through public deliberation on AI, diverse perspectives on bias experienced in the public sector could help inform what characteristics are important for a specific decision context. This could further stretch to a discussion of what fairness metrics or distributive justice principles the citizens would consider appropriate for different situations.

4.5. Fairness beyond decision-points

4.5.1. From formal to substantive equality

The incompatibility between the different mathematical definitions of fairness creates a dilemma called “the impossibility of fairness” (Green, 2021). It speaks to the fact that fairness criteria alone do not necessarily mean the system will have fair outcomes. Since algorithmic fairness restricts the analysis to isolated decision-making procedures, applying a certain metric to make algorithms “fair” can still lead to models that exacerbate oppression. In an unequal society, decisions rooted in formal equality can produce substantive inequality.

Formal equality relates to a liberal egalitarian notion that "when two persons have equal status in at least one normatively relevant respect, they must be treated equally with regard in this respect" (Gosepath, 2021). This means that formal equality restricts its analysis to a single decision point and, therefore, cannot account for the inequalities surrounding that decision point (Green, 2021). To counter this, egalitarian thinkers have tried to expand the formulations of equality to include social relationships and institutional arrangements. This has been named substantive equality, which is oriented towards identifying and remediating social hierarchies. Substantive equality theorists are concerned with abolishing the social conditions that facilitate domination and oppression, as they argue that these social hierarchies are the core problem of unequal distributions of benefits and harms.

Instead of looking at universalist notions of equality, such as equality of opportunity, substantive equality asks that we look at the concrete historical realities that generate injustice and change them (Coeckelbergh, 2022b). This requires a focus on the groups that have historically been marginalised and oppressed, such as women, people of colour, LGTBQIA+ people, indigenous people, disabled people etc. In line with a technological materialist approach, it becomes important to investigate which groups the technologies and those who employ them discriminate against and asks how this happens beyond a single decision point. This requires looking at the larger societal structure of inequality perpetuated by algorithms.

4.5.2. Disproportionate impacts of algorithms

Eubanks (2018) argues that the "new data regime" often makes things more difficult for poor and working-class people instead of empowering or benefitting them. Decision-making about eligibility for benefits, for example, means that poor people are often surveilled and managed, which opens up possibilities for manipulation. If a government has moralistic views of poverty or group membership, the technology can be used to perpetuate biases against already disadvantaged and marginalised people. An example is the child-benefits scandal in the Netherlands, where Dutch tax authorities used a self-learning algorithm to create risk profiles for families committing child-care benefit fraud (Heikkilä, 2022). This resulted in an estimated 26 000 parents being wrongly accused of child-benefit fraud, which required families to pay back the allowances, driving families into severe financial hardship or bankruptcy (Henley, 2021). The procedure was ruled discriminatory and accused of involving racial profiling.

The distribution of benefits and harms of algorithms also relates to questions of access and knowledge of algorithms as well as resources to appeal them. Research has shown that structural inequalities mean that advantaged individuals can exercise greater autonomy in their use of information technologies than disadvantaged individuals (Cotter & Reisdorf, 2020). Furthermore, it has been shown that algorithmic knowledge gaps are greater for people with a worse socio-economic background in certain domains. Even though disadvantaged individuals are more likely to be subjected to harmful algorithmic practices, they are less able to appeal these decisions. Hence, citizens do not have equal capabilities to advance their interests in relation to algorithms. Public deliberation on AI could be a means to adjust for that by inviting citizens to be educated about AI and allowing them to advance their views and experiences. Public deliberation on AI is a

good first step in giving the public the means to become more capable of taking action against unfair algorithms and educating them on how algorithmic fairness might come about.

4.6. Equality and justice conclusion

Algorithms used in the public sector suffer the same vulnerability as any other algorithm, meaning they have the potential to discriminate. This is especially worrying since AI solutions work at scale. As opposed to harm created by a biased human decision-maker in a single decision point, AI risks causing the same harm to a large number of people at once. This means AI risks disrespecting the democratic values of equality and justice.

Ensuring bias is dealt with to prevent harm is especially important in the public sector, as democratic countries usually have equal treatment of citizens as a principle that is enshrined in the constitutions. The introduction of algorithms into society has highlighted how society is already biased, which has opened up avenues to discuss why this bias exists and what can be done about it. Simons & Frankel (2023) argue that it is impossible to take a neutral stance towards the injustice of the past by striving to build more accurate tools. This will only ensure that the unjust patterns of the past get reproduced in the making of the future. Reinforcing the status quo is not a neutral position but a political one implying that past injustices are acceptable for the future.

Measuring the political impact of AI is difficult, which raises questions about how one can evaluate the different operations, practises, interpretations, and perceptions of algorithmic governance (Coeckelbergh, 2022b). One way would be to hold public deliberations about AI in the public sector. This could bring out many interpretations and perceptions about how AI affects the polis and give insights into different sources of algorithmic bias and mistreatment. This could further inform what actions one could take and what would be acceptable to the public.

Discrimination caused by algorithms is not only a question of equality and justice but also about freedom, as injustices impact what people are free or able to do. The different ways AI in the public sector can promote or undermine citizens' freedoms will be discussed next.

5. Freedom

5.1. Definition of freedom

5.1.1. Freedom and democracy

Freedom is one of the most central values and principles in liberal democracies. In the history of political philosophy, the preservation of freedom has been understood as the issue that necessitates political authority and is, therefore, also the most important function of political authority. Hobbes argued that the state of nature, without political authority, would be a state of chaos if humans were allowed to pursue all their wishes without any constraints. If everybody were to exercise their full freedoms, doing as they wished, they would inevitably start interfering with other people's freedoms. Therefore, political authority would be needed to keep a society's internal peace and order, ensuring that citizens exercised their freedoms in a way that did not interfere with others' (Coeckelbergh, 2021). Other philosophers argued for an expanded role of political authority and, therefore, for the role of freedom. For Rousseau, the minimal political authority in a Hobbesian society would not be enough to ensure freedom for people, as it would only benefit those who had property and resources. At the same time, the disadvantaged would remain unfree (Bertram, 2023). He argued instead that humans will only be free if they have been able to have a say in the "general will", which is the collective will of all citizens. This will is the source of law, and since everyone has participated in the formation of this law, they are acting according to their own will and so remain free. For Rousseau, freedom is achieved through collectivity and each individual's participation in exercising control of society. This binds freedom to democratic practice. According to his understanding, a democratic society is free if it is a self-determined society, and each member of society is free to the extent that they participate in the democratic process.

Freedom can be conceived of as both an intrinsic quality of democracy and an outcome of democracy. The fact that each individual in the collective is free to contribute to the direction of the community and therefore decide how they will be governed means that democracy is a form of freedom. Freedom is also an outcome of democracy since when everybody participates equally, it forces decision-makers to consider the interests and rights of a wider range of subjects, thus better advancing those. A robust empirical correlation exists between well-functioning democratic institutions and strong protection of core liberal rights, such as rights to a fair trial, freedom of association, and freedom of expression (Christiano & Sameer, 2022).

5.1.2. Positive and negative freedom

Western political philosophy usually has two conceptions of freedom; negative liberty and positive liberty. Isaiah Berlin defines negative and positive liberty as two contrasting or complementary ways of thinking about freedom (Berlin, 1959). Negative freedom is understood as the absence of external obstacles to the agent. This means that a person is free if no one, whether a person or an institution, is stopping them from doing what they want to do. Negative liberty concerns the options or choices left to the individual to make on their own, what options are not allowed to be

made by the individual, and for what reason (Taylor, 1979). People usually differ in what reasons or values they think legitimise a violation of negative freedom, which gives rise to the concept of positive liberty.

Positive freedom has several definitions. It can be seen as a property of the collective that is achieved through equal participation in the process of establishing self-rule. As such, it is freedom for the collective according to the "general will" described by Rousseau, a freedom of participation. Positive freedom also has an individualist application that addresses individual self-rule, linked to self-mastery, self-determination, self-realisation and autonomy (Carter, 2022). For positive freedom, people are free if they can control their destiny in their interests. This is also sometimes taken to mean that people should be able to control their desires or thoughts in alignment with their interests.

Positive liberty is a response to the perceived inefficiency of the concept of negative liberty to truly enable freedom (Carter, 2022). Theorists endorsing positive freedom argued that there might be many times in an individual's life when no other agent is restricting them, but they are still unable to do what they want. Such conditions could, for example, arise because of economic or social inequalities, such as poverty, or discrimination based on characteristics such as gender, race, or religion. Because of this, positive freedom requires that there are structures in place in society that enable individuals to control their lives so that they do not have to depend on the will of others. The structures for promoting positive freedom aim to give individuals the capabilities, opportunities, and resources to achieve self-realisation, such as free education, free or subventioned health care, or social safety nets.

5.1.3. Freedom and the role of government

Because there are different conceptions of liberty, there are also different conceptions of the task of government (Bertram, 2023). People who endorse positive freedom hold that the government should only interfere if it has a good reason or to promote positive freedoms. The welfare state can be argued to fulfil these criteria as well as universal basic income. In contrast, people who strongly endorse negative liberty argue that the government should only interfere to protect the people's negative liberty. For these people, structures and practices aimed at promoting positive freedom, for example, free health care, could be seen as an attack on negative liberty since free health care would require taxation, which some might argue is an unjustified interference. Negative liberty is commonly used as a defence for constitutional liberties such as freedom of movement, religion, and speech (Carter, 2022).

The public sector's use of AI technologies can impact negative and positive freedom differently. Discerning how it could do so would enable a deliberation on how to balance different kinds of freedoms against other values that could be created or impacted by AI in the public sector. This will be the aim of the following sections.

5.2. Promoting positive liberty with AI

There are many ways in which AI technologies in the public sector could promote positive freedom or influence the conditions for positive freedom, thus giving people the power and resources to self-realise. The algorithms' efficiency could contribute to this, as effective distribution of public goods or services could enable citizens to receive and gain access to what they need or are afforded quicker. Sherover (1992) argues that the positive freedoms afforded to citizens in a society specify what the society encourages citizens to do or not do. This means that AI technologies can make certain activities easier to engage in for citizens. For example, if the government wants to promote economic growth, it could leverage AI to make it easier for individuals to start a business or change careers. The government's use of AI technologies could promote positive freedom by freeing up time, enabling the citizens to plan their lives better and hence be more in control of what they do. For example, a chatbot could be used on governmental websites, guiding citizens and helping them navigate to the services and information they need, ensuring citizens do not have to go to specific locations for information (Miller, B., 2021). Another example could be the use of algorithms to analyse and manage traffic. Computer vision systems can analyse data of real-time traffic flows and send that information back to citizens to help them plan their routes, enabling ease of movement and helping them save time (Sajid, 2023).

If the government provides healthcare, AI technologies in healthcare could promote positive freedom since any sickness or illness could hinder positive freedom. For example, AI can predict potential illnesses from patient data, facilitating early prevention methods and potentially saving patients from many burdens. AI can also detect and control a virus outbreak, as was done during the COVID-19 pandemic (Sajid, 2023). An example is the government's PHREDSS system in Australia, a surveillance system that monitors symptoms in hospitals daily to detect emerging disease outbreaks quickly and change health policies proportionately (Centre for Epidemiology and Evidence, 2022). Such a system could also promote negative liberty, as full outbreaks might restrict citizens' movement, which would violate freedom of movement, hence negative liberty. AI technologies could also be used to analyse public data, allowing policymakers to identify what issues need to be addressed in different regions allowing them to respond quicker, hence using data to distribute resources better.

The government's use of AI to track and prevent cyber attacks on its citizens is another way of promoting positive and negative liberty (Fadia et al., 2020). If a cyberattack leads to the internet being inaccessible for a few hours or more, citizens' lives would be negatively impacted as they would not have access to resources and capabilities that they normally do, therefore being less able to control their lives. A cyber attack would be a violation of both negative and positive liberty.

There are many challenges with implementing AI technologies in the public sector, which could instead contribute to violations or reductions of positive freedom. These will be discussed next.

5.3. Obstructing positive liberty with AI

5.3.1. Bias and categorisation

If positive liberty is concerned with the structures that aid people in realising their goals in life, then any obstruction to these structures could be seen as a reduction in positive liberty. This means that if public resources become unavailable to the people who need them, it would reduce their ability to control their lives and diminish their positive freedom. This could happen due to discrimination. An example is the child-benefit scandal in the Netherlands which was shown to be based on racial profiling. Beyond algorithmic bias, if an algorithm is not robust enough in its problem conceptualisation and specification, it may lead to lower accuracy and to arbitrary results, which could have the same effect of reducing people's abilities to access services and goods that they have a right to.

Many services the public sector provides depend on people fitting into different categories that determine eligibility for a specific service or benefit. This categorisation can happen due to the extensive datafication of citizens by the state, which reduces citizens into a flow of data that the "user", here the government, can easily control (Borgmann, 1992). This reduction of citizens into separate categories can also be seen as a representation of what categories are considered "normal" at a certain time. Data is an integral part of the modern state, as statistical information is necessary for any democracy and is used both to dictate the state's duty and to measure its success (Desrosières, 1998). Data about citizens, for example national censuses, have been collected for hundreds of years. In this history, there are many ways that the creation, deletion, and wording of bureaucratic categories have effectively erased many people's identities, histories, and demographics (Ananny, 2021). Who is accepted and who is not, and what is normal and what is not, is therefore linked to the beliefs, interests, and goals of the current norms of society, influenced in many ways by those currently governing.

5.3.2. The example of automated gender recognition

The categorisation used for algorithmic systems might mean that some people's identities are reduced, misrepresented and or eliminated if they do not fit into the categories considered normal. An example is how demographic gender categories used in many algorithmic systems are often a strict binary of male and female (Scheuerman et al., 2019). This excludes trans and non-binary people who often get misclassified. Efforts have been made to be more gender-inclusive in governments, including online forms that allow people to self-identify (Leufer, 2023). However, this choice is not available when it comes to automated technologies, such as AI. The drive to classify people along the lines of gender has given rise to the use of facial recognition systems for the task, so-called automated gender recognition (AGR).

As AGR systems usually operate along a gender binary definition of gender, they consequently harm trans and non-binary people. Trans people are harmed as they risk being misgendered by the system. Non-binary people are harmed as they risk not being classified by the algorithms, which could also create problems. If these technologies were integrated into spaces considered gendered, such as bathrooms, they could cause significant harm to these groups of people. Gendered

bathrooms are normally a difficult space for trans and non-binary people as it requires them to choose one space over another, leaving them at risk of violence or threats from cis users of the facilities (Keyes, 2018). If AGR systems were deployed to monitor these spaces, it could mean that a system could raise an alarm when ambiguous or "incorrectly" gendered subjects entered a gendered bathroom. Some literature proposes using operators to handle these situations (Santana et al., 2017). This operator could ID the person and force them to use the bathroom of the gender they were assigned at birth. Alternatively, they could force them to leave. They could also potentially alert the police, which would further endanger the person, as police discrimination and violence against trans people is a known problem (Grant et al., 2011; Miles-Johnson, 2015).

If systems like AGR were adopted by the public sector and used to monitor bathrooms or other gendered spaces in government buildings, hospitals, or vaccination centres for example, it could exclude trans and non-binary people from these spaces (Leufer, 2023). This would be a violation of their positive freedom, as trans and non-binary people might then be discouraged from accessing the services that they have a right to, as well as being dehumanising since it would prevent these people from performing their basic needs. Such systems would also violate their negative freedom, as they would not have freedom of movement in public spaces.

5.3.3. The price of not being recognised

As seen in the previous discussion, positive liberty is highly related to questions of equality and justice. Philosopher Axel Honneth (2001) has argued that struggles over distributions (benefits and harms) are closely connected to struggles for recognition. He writes that the struggle for recognition "represents a conflict over the institutionalised hierarchy of values that govern which social groups, based on their status and esteem, have legitimate claim to a particular amount of material goods. In short, it is a struggle over the cultural definition of what it is that renders an activity socially necessary and valuable." (p.54). If the algorithms do not recognise certain groups of people, they risk harm in the form of discrimination but also risk not having their needs met by the government in other ways. As mentioned before, governments use data on citizens from different data sources to reason about what public issues need to be addressed. Any data mining that relied on scraped data from the internet to create profiles on people with binary gender demographic categories would risk erasing important understandings about the lives of trans and non-binary people and their specific needs. If data along gender binary categories were used to reason about who needs better services and what services they need, then the needs of those not fitting those categories would be left out. This is another way that algorithms used in the public sector could reduce or violate people's positive freedoms.

The question then is what to do about these challenges since AI technologies rely on categorisation. One could argue that some technologies, such as AGR, should be banned since they discriminate based on gender. Furthermore, even though more costly, data gathering on the experiences and needs of marginalised individuals could be done in more traditional ways, complementing or substituting algorithms for the job. There are some efforts to create gender-inclusive datasets for machine learning. However, since gender is a fluid concept, classifying gender is difficult, and the idea that gender can be classified can itself be critiqued (Pareek, 2019).

Inviting those vulnerable to algorithmic harm to participate in design processes and deliberations about impacts and visions for algorithmic use is important so that algorithms in the public sector could facilitate emancipation rather than erasure, harm or neglect. However, the onus must be on those who deploy the systems to prove that the algorithms are safe and do not contain bias rather than putting the burden of proving that is not so on vulnerable individuals. Goodin (1985) writes that "failing to take positive action to prevent harm from befalling someone who is particularly vulnerable to your actions and choices is morally akin to a bodyguard sleeping on the job" (p.111). This means that systems need to be weighed in terms of their benefits and harms but that the main focus should be on ensuring that these systems work well for all citizens while enabling a plurality of self-identities. Positive freedom should be promoted for all rather than only those fitting neatly into the currently normalised categories.

Coeckelbergh (2021) argues that citizens can demand that any use of AI preserve negative liberty and promote positive liberty. One technology used by the government which by some can be argued to do both and by some is argued to violate both, is surveillance, which will be discussed next.

5.4. Surveillance - protecting or violating negative liberty?

5.4.1. Current surveillance practices

Surveillance is the monitoring of behaviour, activities, or information with the goal of information gathering, influencing, managing or directing said behaviour or activities (Lyon, 2001). These activities can be offline, such as driving on the motorway or grocery shopping, or online, such as websites we visit or friends we are connected to on social media. Whereas previously, surveillance was mainly done by having humans perform the analysis, such as analysing CCTV footage for identification of suspicious activity, the combination of AI technologies and big data now enables the automated monitoring of the surveillance systems, thus reducing the costs and time constraints presented by human workers while simultaneously enabling a more ubiquitous and pervasive role of surveillance in society (Viola, & Laidler, 2022).

Governments worldwide employ surveillance to ensure public safety and national security (Viola & Laidler, 2022). The state uses surveillance to enable law enforcement and intelligence agencies to collect data in domestic policing and counter-terrorism investigations. AI-based surveillance activities are usually twofold: identifying individuals through different sensors or processing data to gain intelligence. The identification of individuals includes, for example, using CCTV footage collected in public spaces or body cams on police officers. Facial recognition systems are being used at airports and borders, which enables the state to identify and track the activity of flagged or suspicious individuals based on facial patterns (Coeckelbergh, 2022b). Other biometrics that can be analysed with AI technologies and used for identification are fingerprints, a person's gait (way of walking), or voice (Sien et al., 2019; Ali et al., 2021).

For prediction and intelligence gathering, social networks such as Facebook can be analysed for cyber threats by detecting communities of interest, identifying network leaders and experts, and analysing text in and between networks (Kirichenko et al., 2018). Another way that the government can surveil its citizens is by creating data profiles on them from online and offline recorded activities. This is done through data mining techniques, which use statistical techniques and algorithms to find patterns of correlations between data (Hildebrandt, 2008). This can be used to identify and represent individuals or groups as data subjects for risk assessment or assessment of opportunities for the subjects.

5.4.2. The question of interference

Surveillance can be seen as an effort to protect citizens' negative liberty by ensuring that nobody negatively interferes with their lives by hindering terrorism or preventing cyberattacks, for example. It could also be seen as protecting people's positive freedom, preventing them from harm and ensuring that they are in control of their own lives. However, surveillance could also be understood as an interference by the government in people's lives, violating negative liberty, as it includes monitoring and watching people's actions. Some might oppose such claims since watching does not necessarily mean interfering. Mostly, the state is not actively interfering with a person's life physically. When they do, they might only do so to protect other people's freedoms or because of other values that might take precedence in the situation. A problem arises when these technologies produce errors that lead to discriminatory and harmful uses such as false accusations and arrests, something that has already happened in the US when a man was falsely accused of stealing thousands of dollars of watches by an error in the facial recognition software and consequently arrested (Bhuiyan, 2023).

Some might argue that for surveillance technologies, even the risk of interference constitutes interference, as the mere threat of interference is enough to discipline you (Coeckelbergh, 2022b). A common analogy for the disciplining power of surveillance technology is the panopticon. The panopticon was a prison architecture designed by English philosopher Jeremy Bentham in the 18th century which was designed to make the prisoners feel like they were always being watched. Bentham argued that this would force the prisoners to behave by design, creating the desired effect while minimising the need for prison guards (Zappe & Gross, 2019). In today's society, these ever-watchful eyes are now non-human, constantly present in our devices, surroundings, and awareness. As for the panopticon, people cannot watch them back.

5.4.3. Privacy and surveillance

With surveillance, what can be seen as being violated in terms of negative freedom is privacy, which can be understood as freedom from interference in our personal sphere (Coeckelbergh, 2022b). Privacy is a human right determined by Article 12, which states that: "No one shall be subjected to arbitrary interference with his privacy, family, home or correspondence, nor to attacks upon his honour and reputation. Everyone has the right to the protection of the law against such interference or attacks." (UN, 1948). Surveillance can harm through its forceful entering of spaces which were previously only privy to one's self, such as one's questions and wonders (through, for

example, one's Google searches), one's correspondences (for example, one's online messages and emails), one's choices of movement (through CCTV footage in streets), and in the future perhaps even one's brain patterns (through neurotechnology). Like the panopticon, it can have a disciplining effect that changes one's behaviour due to intimidation and shame that could come with the knowledge that one is being watched (Couldry & Mejias, 2019). In this way, surveillance is oppressive and productive, creating the individual according to the norms of the dominant culture and shaping one's identity by "training us to look at ourselves in certain ways" (Zappe & Gross, 2019, p.14). If unfreedom from a negative liberty perspective is the restriction of actions or outcomes that a person would otherwise be able to do or bring about (Carter, 2022), then the disciplining effect of surveillance could be seen as a violation of one's negative freedom if it results in a restriction in what one says (freedom of speech), in where one goes (freedom of movement), and in whom one meets (freedom of association and assembly).

However, the question is whether surveillance poses an 'arbitrary interference' in citizens' lives and what degree of interference it presents. Is the degree of interference serious enough to violate negative liberty? Some have argued 'no' based on people's attitudes towards surveillance and claim that the idea of surveillance as a disciplining force is no longer valid (Zappe & Gross, 2019). They claim that people nowadays are aware of being watched and that despite knowing this and without knowing how they are being watched, they do not regulate their behaviour. Instead, as Vaidhyanathan (2012) writes, "we don't seem to care" (p.85). The act of surveilling is not only done in a top-down fashion, from the state to the citizens, but also throughout society and between people. Hence, the practice of surveillance might be a social phenomenon that people are acclimated to. As Foucault (1995) observed: there are multiple "centres of observation disseminated throughout society" (p.208).

People monitor each other on social media, scolding bad behaviour and encouraging what they perceive to be correct according to norms. Self-monitoring is a common practice for both institutions and individuals, and people both seem to fear privacy invasions but are simultaneously unbothered by sharing information about themselves through social media (Harju, 2019). Some might even claim that surveillance might not be seen as such if the monitored subjects freely share their information. Zappe (2019) writes that distinguishing between public and private, and data and the "self" is becoming increasingly difficult. Therefore, there might come a time when the state's role as a disciplinary institution is supplemented by people's "over-sharing, consumerism, and indifference" (Zappe & Gross, 2019, p.15).

Surveillance used by the public sector to ensure public safety and order may not be seen as a violation of negative liberty either because it does not interfere or because, even if it does, it does not hinder people from doing what they would normally do. However, some might argue that the further actions enabled by surveillance combined with other AI technologies pose a greater risk to citizens' freedoms. These will be discussed next.

5.5. Controlling citizens with AI

Governments can use the power that comes with surveillance to control the citizen and direct them towards certain aims. This relates to the practices of paternalism and authoritarianism. This section will first discuss how AI can be put to authoritarian use. After that, I will lay the theoretical ground for paternalism and show how it relates to the practice of nudging. I will then show how AI technologies can be used for nudging and discuss the implications of such technologies on freedom.

5.5.1. Authoritarian algorithmic practices

With surveillance, the government can access many of the movements and actions of citizens. If these actions are accessible, then with the right technology, the government could also try to influence these actions. Hence, with AI technologies, including automated surveillance, the government could perform mass manipulation and repression. Such a tool in the hands of a government that is currently or is becoming totalitarian or authoritarian could mean that many of citizens' liberties would cease to exist or be severely reduced. Surveillance could be used to detect any form of speech by individuals or groups that do not align with the regime's ideology, thereby identifying potential political opponents and enabling them to take action against them. This would clearly violate negative freedom, as freedom of speech, assembly, association, and religion or beliefs could be impacted. Paired with other technologies like generative AI, surveillance and targeted action makes it easy for authoritarian regimes to create and distribute propaganda. An example is Vladimir Putin's use of AI-enabled propaganda to manage Russians' views and feelings on the current Ukraine invasion. Generative AI was for example used to create a fake video of Ukrainian President Volodymyr Zelensky in which he calls for Ukrainians to surrender (Drexel & Withers, 2023). These examples show that AI technologies can be used in inherently anti-democratic ways.

5.5.2. Paternalism and nudging

Democratic governments might also interfere with citizens to achieve their interests. To reduce public healthcare spending, the government could take measures to ensure citizens' good health. A non-AI example is how Sweden has banned cigarette smoking at bus stops, train platforms, outdoor seating areas, and outside entrances of hospitals and other public buildings to reduce smoking and its health hazards (CBS News, 2023). This is a form of paternalism, which is interference against somebody's will motivated by the claim that the person would be better off or protected for it (Dworkin, 2020). Paternalism involves the restriction of some freedoms to protect or promote other freedoms and values.

Government interference in people's lives is often done in the name of positive freedom. This requires a specification of what should be encouraged, hence the content and activity of freedom (Bowring, 2015). Because of these constraints, Berlin mounted a famous critique of positive freedom by showing how it could sometimes lead to paternalism by those who claim they know our true interests better than ourselves. In his argument, the premise for this kind of positive freedom begins with the idea of a divided self (Carter, 2022). This could be the difference between

one's rational self and one's desires, one's short-term thinking versus long-term thinking, or other theories that distinguish between different ways of thinking, feeling, or being that can represent a conflict in one's self or the greater collective (as an extension of self).

One of the current popular theories that can motivate paternalism in this way is Kahneman's two-system theory of the mind. This theory stems from work that Kahneman did early in his career with Tversky on human decision-making (Mheslinga, 2022). Their experiments showed that humans do not commonly reason in line with statistical and logical thinking in situations of uncertainty, instead relying on different heuristics for their judgements. They used this evidence to argue that humans are irrational decision-makers (Tversky & Kahneman, 1974). In the two-system theory of the mind by Kahneman, he describes the brain as a lazy machine that keeps people from using the full power of their intelligence (Kahneman, 2011). Instead, most people's behaviours are determined by two systems, one that is conscious and considerate, which is used for focused reasoning, and one that is automatic and impulsive, which uses heuristics to solve problems fast. According to Kahneman, these two systems are constantly competing, and different external situations can trigger one system over the other.

The studies by Kahneman and Tversky and Kahneman's two-system theory have been criticised on several grounds, for example for lack of replicability and wrongful conclusions from the results (Gigerenzer, 1991; Schimmack et al., 2017). Despite this, the theory that humans often use heuristics to make decisions served as an influence for the development of nudge theory by Thaler & Sunstein (2008).

Nudging is the idea that environments should be designed to influence the decisions or behaviours of groups of individuals in a desirable direction, knowing that people will often not reason well about their choices on their own. An example is putting fruit next to the supermarket check-out instead of candy to reduce people's sugar consumption. Nudging is seen as an alternative to coercion, as it makes some choices more available than others, leaving it to the individual to make the final choice. Sunstein & Thaler call this "libertarian paternalism" and mean that nudging does not violate (negative) liberty like "classical" paternalism, as the choice is still up to the individual.

By steering people towards "better" choices that are aligned with people's or the collective's true interests without forcing them to make those choices, nudging can be said to promote people's positive freedom while preserving negative liberty. However, this is the kind of action that Berlin warned against. He meant that manipulation towards goals that the social reformer recognised but the people do not is degrading, as it treats them as objects without wills of their own (Berlin, 1959). To Berlin, nudging would violate one's negative freedom as it tries to influence people's choices when they might have acted otherwise.

Regardless of this critique, governments worldwide have been enticed by nudging as a tool for directing the behaviours of their citizens in a less interfering way (Halpern & Sanders, 2017). Nudges have been used to get people to save for retirement, eat more healthily, and pay their taxes on time. Some have also started thinking about how governments could pair AI technologies with nudging to create nudges that work on scale and are more targeted for individuals, functioning as

regulatory technology (O'Reilly, 2013). Cristianini & Scantamburlo (2020) have investigated how AI and nudging could be paired for purposes of social regulation. They are interested in the implementation of what they call Algorithmic Social Machines (ASMs) for social control, which would work like personalisation systems, recommending and discouraging actions of citizens in real-time through surveillance and administering positive and negative incentives (such as fines or discounts) to influence their behaviours. An example of a small-scale system of this sort is a smart traffic app launched in Enschede in The Netherlands which rewards people for cycling, walking, and using public transport (Huang et al., 2021). The rewards consist of points that can be redeemed towards discounted products and services in the local community. Likewise, the social credit system in China follows the same logic, using AI technologies to monitor, analyse, and then steer people's behaviour by rewards and punishments (Lam, 2021).

5.5.3. Justifications for nudging

In a democratic system, it might be possible to justify such a system based on positive freedom. If positive liberty concerns agreements made together as a community that all people have participated in (according to Rousseau's general will argument), and if such an agreement involves paternalistic supervision in situations where people's competencies can be doubted, then paternalism would not be any infliction on liberty despite people being handled by someone else. To illustrate this, Reploge (1989) uses the metaphor of somebody taking a person's car keys against their will after drinking. The problem with algorithmically superpowered nudging is that this kind of agreement is lacking, or can be said to be lacking since there has been no public involvement about the issue that determines what the goals of such a system should be. If an algorithmic nudging system was used to encourage saving up for one's retirement, this might be seen as more benign than a nudging system for managing people's movements for example. However, a system that exploits citizens' subconscious would have to be accepted and decided on by the citizens. If the system deals with punishments and rewards, it would have to have the goals of the system explicitly stated. Such a system would be more transparent and might paradoxically be seen as less liberty-inflicting than systems used without people's knowledge.

Despite only managing choice architectures, it is possible to ask how liberating an algorithmic nudging system would be, especially when rewards and punishments are involved. Instead of supporting the citizens, it can be argued that this system does the opposite since it blames individuals for their shortcomings or bad luck. As many political problems are wicked, pertaining to different needs, values, and wishes for difficult situations, people might feel differently about the uses of such a system given its potential impact on their freedoms. The impact would likely differ for different groups of people, further necessitating their involvement. Because such a system could easily be misused in anti-democratic ways, it would be important that the citizens are aware of algorithmic nudging systems and could contribute to the goal-setting and envisioning of its possible uses. Even though some goals might be indisputable and therefore would not necessitate public discussion, such as the goal to save the planet from the climate crisis, involving the citizens in envisioning exactly how this could be done would be important from a deliberative democratic perspective. Public deliberation on the government's uses of algorithmic nudging could contribute

to a discussion of what values need to be balanced and considered for such a system, how potential risks can be mitigated, and what uses would be most effective and acceptable to the public.

5.6. Freedom conclusion

Freedom is one of the most central principles in liberal democracies and also one of the most complex. As shown in this chapter, introducing AI technologies in the public sector raises important questions about how the technology might impact this principle. AI in the public sector holds the potential to make people's lives easier by facilitating better planning, ease of movement, fast acquisition of resources, and safety from external threats. This means that AI has the potential to promote the positive freedom of citizens, enabling them to better plan and be in control of their lives. However, as Berlin has argued, references to positive freedoms can also be used to justify forms of paternalism, which might impact citizens' negative freedom by interfering (or potentially not) in their everyday lives. Furthermore, surveillance technologies could affect people's privacy, raising important questions on how to protect this principle while reaping other benefits associated with surveillance. It is up to the public sector to provide information about the risks of the harms that can be done by these technologies alongside the expected benefits so that the public can assess whether they agree with the priorities set.

In a public deliberation on AI use in the public sector, citizens could give reasons as to what restrictions of freedom in reference to another principle would be acceptable and which ones would not. This could further establish what new kinds of freedoms could be created to compensate for the loss of others. Verbeek (2020) writes that since opting out of technology is not always possible, we must look for a democratic perspective from "within". This means acknowledging that individual freedom is always mediated and conditioned by technologies and choosing to accept responsibility for and deal with this fact in a democratic way. Public deliberation would be a democratic way of discerning how AI could make the government more efficient while still protecting and promoting citizens' freedoms.

As I have shown, AI in the public sector impacts the values of freedom, equality, and justice. Because of this, it is important to ask whether the use of AI in the public sector is legitimate and whether the citizens could hold the government accountable for their AI use. These issues will be discussed next.

6. Legitimacy

6.1. Definition of political legitimacy

Political legitimacy can be minimally defined as the right to govern (Coicaud, 2002). It is a virtue of both political institutions and the decisions made within them (Peter, 2017). For many contemporary political philosophers, the right to govern is dependent on a democratic system, where those who are being ruled over get to vote and participate in the governance of society. Theories of political legitimacy can be divided into three categories: those that focus purely on the quality of the outcomes of political processes for the claim of legitimacy (instrumentalists), those who mean that it is the quality of the decision-making procedures that gives a decision or a political institution legitimacy (proceduralists), or a mixed approach of the two (mixed instrumentalist and proceduralist). In order to avoid mixing instrumentalism in political legitimacy with instrumentalism in the philosophy of technology, I will refer to it here as *the outcome approach*.

The outcome approach believes an institution or decision is legitimate if it leads to good or ideal outcomes. "Ideal" here can refer to an ideal egalitarian distribution (Peter, 2017). People subscribing to the outcome approach can sometimes argue against democracy if it does not lead to good or ideal outcomes. For the legitimacy of decisions, they might argue that experts should make certain decisions if it leads to better outcomes rather than relying on democratic decision-making. Because of this claim, some scholars have tried to show that a democratic system leads to better outcomes than other systems. For example, Landemore (2013) uses the diversity trumps ability theorem to argue that democratic decision-making leads to good outcomes since a diverse group of people coming together will find the best outcome over a group of experts.

In contrast to the outcome approach, pure proceduralists argue that political institutions and their decisions are legitimate if they result from fair or correct procedures. Fair here refers to political equality, which for the aggregative democratic model demands that all expressed preferences are given equal consideration. For the deliberative democratic model it demands people's equal possibilities to participate in the process of public deliberation (Peter, 2007). Deliberative democracy can be seen as a proceduralist account since, in deliberation, there are no shared standards for assessing the quality of the outcomes. Deliberative procedures always have disagreements about the proposals remaining. Therefore, some deliberative theorists argue that one can only ever say something about the fairness and quality of the procedure.

Against these two theories of political legitimacy, proponents of a mixed approach hold that the fairness of a procedure alone does not guarantee good outcomes. For them, fair procedures might still lead to irrational outcomes or outcomes of unacceptably low quality (Peter, 2017). There are different conceptions of mixed approaches. Deliberative democrats that agree with the mixed approach argue that the structure of deliberative procedures will eventually lead to rational outcomes (Habermas, 1996). Other deliberative democratic theorists have tried to prove that there is still the risk of irrational decisions from a deliberative procedure by devising different dilemmas and proving them. Many have therefore turned to an epistemic conception of proceduralism,

which holds that deliberative democratic decision-making procedures are required for political legitimacy but also require the decisions made to be as correct as possible.

6.2. The principle of publicity

In order to evaluate the legitimacy of a government or decision, a kind of openness is required between the government and the people. This relates to the principle of publicity. Publicity requires that political decisions that can affect the happiness or welfare of citizens should be based on reasons available for scrutiny (Davis, 1991). Hayward (2021) writes that the principle of publicity constitutes at least three values. The first is openness or inclusiveness, which requires that spaces, institutions, or goods considered public should be accessible and open to all, hence being of common concern. The second is transparency. Transparency is necessary since the public should be able to know how public things are managed and governed. This includes the notion that reasons behind decisions made about citizens should be accessible. The third value of publicity is its politicising character, which emerges as public things encourage people to understand themselves as political actors and enable them to participate in caring for those things.

The function of publicity is to foster openness and trust between those governing and the governed by enabling the citizens to manage power relations democratically. Hence, the principle of publicity is required for the citizens to assess the legitimacy of democratic institutions and their decisions. Considering these conceptions of legitimacy and the associated principle of publicity, we can begin to consider how AI technologies in the public sector might impact democratic legitimacy.

6.3. How AI in the public sector can affect legitimacy

Political legitimacy is a virtue of both institutions and decisions, which means we can analyse AI in the public sector from different points of view. When looking at institutions, we could analyse whether an institution that uses AI for most of its decisions rather than humans would be legitimate. We can call this *algocratic legitimacy*. In terms of the legitimacy of decisions with AI, there are two points of analysis: 1) whether the decision to apply an AI solution to a problem is legitimate, and 2) whether the decision by an algorithm is legitimate (here we assume that the prediction task and the decision task is the same). We can call these *AI solution legitimacy* and *algorithmic decision legitimacy,* respectively. I will discuss these different conceptions of AI legitimacy by examining how they measure up to the proceduralist and outcome-approach, respectively.

6.3.1. The proceduralist account

6.3.1.1. Algorithmic decision legitimacy

From a proceduralist point of view, algorithmic decisions are legitimate as long as the procedures are fair. For a procedural account of algorithmic decision legitimacy, it becomes impossible to speak about fairness from a deliberative point of view since there is no space for deliberative participation inside an algorithm. However, from the aggregative democratic point of view, we can demand that everybody be treated as political equals and that their interests are considered equally. Some might

argue that the fact that algorithms process everybody in the same fashion (using the same internal functions) might indicate that the process is fair. This is in comparison to a human-led process where the human in question might treat different people in different ways due to social bias, such as stereotyping people, leading to everybody's interests not being equally advanced. Hence, algorithmic decisions might be considered procedurally more fair (and hence legitimate) than the same decisions being taken by public servants.

However, a problem arises with the use of data. Even though there is no bias on the part of the algorithm's internal processing, the data is still treated in the process, and from a broader understanding of AI as not just a technology but also as a system, the data can be argued to be part of the procedure of algorithmic decisions. Suppose for example that the data is not representative of the wider population. In that case, everybody's interests will not be advanced equally in the algorithmic decisions, which could make algorithmic decisions illegitimate according to the proceduralist account. Another question is if individuals are treated as political equals or whether their interests instead are subjected to their group membership. AI technologies are often based on statistical similarity, meaning decisions are not made based on the individual and their characteristics but rather on that individual's similarity to other people. This might mean an individual's interests are not advanced if algorithmic decisions are biased against their group.

Here, some might argue that a decision by a human decision-maker might be more procedurally fair since a human assessment could include an understanding of the individual based on feelings of empathy, potentially leading to the advancement of that individual's particular interests beyond judgements of their similarity to other people. To counter that argument, one might argue that if fairness metrics are involved in the process, meaning biases can be caught and corrected, the potential for biases is acceptable, making algorithmic decisions legitimate. However, the question of what fairness metric to apply to an algorithm has to do with decisions to use an algorithmic solution, which will be discussed next.

6.3.1.2. AI solution legitimacy

The decisions to apply an algorithmic solution and the design of such a system can be considered part of AI solution legitimacy. The proceduralist account of deliberative democracy can apply here, meaning such decisions must be subject to public deliberation. Deliberation about such decisions might happen within different institutions and within different design teams. However, according to deliberative democracy, certain issues should be the subject of public deliberation, where the citizens also participate.

Solomon & Abelson (2012) have identified four policy issue characteristics that make it suited for public deliberation. Accordingly, a topic is suited for public deliberation if it has one or more of the following: conflicting public values, high controversy, could benefit from combined expert and citizen knowledge, and low trust in government. These are all fulfilled by AI in the public sector. In the previous chapters, I have shown that how AI is or can be used in the public sector could mean conflicting public values (for example the value of public safety versus privacy). AI is also a topic of high controversy, pertaining to questions of whether AI can be properly controlled and aligned with human values and whether the knowledge created by AI is superior to human knowledge.

These ideas will be further developed in Chapter 7. I have also shown how AI both requires expert knowledge in terms of understanding the algorithms as technical artefacts (instrumentalist view) but also pertains to questions of, for example, what principles of distributive justice and potential fairness metrics would be appropriate in different situations and what specific uses and their resulting social structures could be deemed acceptable. From a deliberative democratic view, the citizens could be involved in deciding this. I have also mentioned empirical evidence that citizens are concerned about governments' application of AI, showing that AI in the public sector is an issue where the public might have low trust in the government.

Because of these factors, in the proceduralist view, the decision to apply AI in the public sector is only legitimate if it is the subject of public deliberation. This is also the main argument I am advancing in this thesis.

6.3.1.3. Algocratic legitimacy

From a proceduralist view, algocracy is a potential threat of the outcome-approach applied to AI. Algocracy can be defined as a system "organised on the basis of algorithms which structure constrain the opportunities for human interaction with that system" (Danaher, 2016, p.251). Danaher (2016) argues that a problem with the outcome-approach to legitimacy is it opens up for an algorithmic version of epistocracy. Epistocracy means that a group of people in society know better what the best outcomes are (an epistemic elite) and, based on this superiority, get to control all public decision-making processes (Estlund, 2003). The argument is that if one believes that only the outcome matters, one should accept such a regime, which would prohibit any citizen involvement in public decision-making procedures. Danaher extends the idea of epistocracy to algorithms. If some believe that algorithms have a superior epistemic access to the best outcomes, this might lead to algocracy. He calls this the "threat of algocracy" since it would diminish human participation in decision-making processes that humans previously controlled.

An algocracy would not be legitimate from a proceduralist point of view unless it results from a fair decision procedure (AI solution legitimacy). However, from a deliberative democratic point of view, it is unlikely that an algocracy would be deemed legitimate since it does not allow for human participation in decision-making. For the aggregative democratic model, algogracy would be legitimate if it equally advances the citizens' interests.

6.3.2. The outcome-approach

6.3.2.1. Algorithmic decision legitimacy

For the outcome-approach, algorithmic decisions are legitimate if they bring good outcomes. This brings the question what the outcome to be evaluated is for an algorithmic decision. If taking an instrumentalist approach, the outcome to be analysed is how well the algorithm has captured reality as represented in the data, hence the model it creates. The outcome of an algorithmic decision could extend to how well such a model performs in the environment it was made for, hence the model's predictive ability. Good from the instrumentalist point of view might therefore refer to an algorithm's epistemic capacity. Epistemic capacity, adapted from *community* epistemic

capacity, can be defined as the capacity of an algorithm to gain, maintain, adapt, and continue the knowledge it needs to solve a problem (Werkheiser, 2016).

One can argue that good algorithmic decisions require good outcomes not just for a single decision but also over time. This could cause a problem for outcome-based algorithmic decision legitimacy, as many current systems are considered fragile. Data-driven algorithms are not robust in their predictions. This is because they rely on underlying statistical patterns in the data to make decisions. As data is continuously updated based on the changes happening in the world, the models might eventually not accurately represent the world, making them prone to errors. Such changes in the data can further be hard to predict, as things might change quickly. Likewise, algorithms might prove fragile as they are based on correlations and not necessarily causations. Thus, according to the data that the algorithm was trained on, the ground truth might not be consistent with the ground truth as it exists in the world.

These fragilities of the systems also mean that algorithms are vulnerable to adversarial attacks, where small and carefully crafted perturbations in the input data cause the algorithms to give false predictions (Wang et al., 2023). Such changes are observable to the human eye and would not pose a problem for a human decision-maker but would cause the AI technologies to fail. Furthermore, with the current rise of generative AI, it is possible to generate life-like images and videos alongside text which looks like a human has written it. These fake media can be used to trick AI systems. Such risks pose further problems than adversarial attacks by perturbations, as not even humans can detect them as fake. This means that despite depicting the world well at the moment of evaluation, algorithms might not do so consistently, which could mean that algorithmic decisions are not legitimate from an outcome-approach.

6.3.2.2. AI solution legitimacy

Considering the decisions to apply an AI solution, decisions are legitimate for the outcome-approach if they lead to good outcomes. Here we can apply both a narrow and broad analysis of the outcome. In a narrow analysis, it makes sense to judge the outcome of the decision to apply an AI solution based on how well it solves the particular problem the institution meant it to solve. Hence, AI technologies used to enforce public policy should be assessed on how well they manage to pursue the aims of the democratic lawmaker. In a representational democracy, that lawmaker has been decided on by voting, and for the outcome-approach, one can argue that as long as the AI serves the ends of that lawmaker in a sufficiently good way it would be legitimate. Whether an AI solves a problem better than, for example, a group of human decision-makers might depend on what evaluation measure is used. Some might for example argue that if the algorithms solve the problem more efficiently than human decision-makers, this might be considered a better outcome. Others might define good in a way that requires the AI solution to solve the problem in a way that leads to an ideal distribution.

The question then is what outcomes an AI solution should be evaluated on and if it is possible to claim the legitimacy of AI solutions based only on its success according to one metric or outcome. Furthermore, whether one could sufficiently separate the outcome from the procedure might be questionable. If an algorithm required a data gathering procedure which treated citizens

inhumanely, for example by collecting images of them in their private spheres, would this be warranted based on any good outcomes of that data gathering, or would the potential harm to the affected individuals also count as part of the outcome (Danaher, 2016)? Mesquita (2019) writes that we cannot divide the world into "a neat dualism of aims and tools - the aims of public policy, on the one hand, and objective, quantitative tools used to pursue those goals, on the other". He means that there is inherent feedback between the two since what we quantify and how we quantify something decide the aims of public policy, as the aims of public policy decide what we quantify. This means that AI solutions might be applied in a way that requires a definition of public policies into measurements that can be made. From a technological materialist analysis, this could mean that outcomes of AI solutions on a systemic level are not considered in the evaluation of the technology. An example is how poor people might be surveilled by the system more than more privileged people, which might cause certain problems that are not factored into the outcome. This means the outcome-approaches to AI solution legitimacy depend on what outcome is deemed important, which might be a different answer for the government and the citizens.

6.3.2.3. Algocractic legitimacy

As discussed for the proceduralist account, the outcome-approach would deem algocracy legitimate if it leads to better outcomes. As discussed above, this depends on what criteria are chosen for what makes a good outcome. A further potential for algocracy is that, whereas for AI solution legitimacy, the outcome to be analysed is selected by humans, for an advanced algocracy, the outcome to be measured could be selected by the algorithmic system itself if given enough agency. However, whether one can even speak about political legitimacy if it is not judged by humans is questionable.

6.4. AI publicity and accountability

Both the proceduralist and outcome-approach depend on the procedure and outcome of the algorithms and the decision to apply them being available so that citizens can judge whether the procedures are fair or the outcomes are sufficiently good. Discerning whether a decision was fair and its outcome good relates to the principle of publicity.

6.4.1. Proceduralist requirements of AI publicity

From a proceduralist point of view, to assess the legitimacy of algorithmic decisions would require that citizens have access to the internal workings of the algorithms and are given reasons for why a decision was taken for a specific case. This is where a potential problem arises for AI technologies: for a human decision-maker, it is possible to ask them to give reasons for their decisions to justify them. This is not possible when it comes to algorithms. There is no logical pathway from an input in a data-driven algorithm to its output that can be read from the code (Maclure, 2021). Some algorithms find very complex relationships in the data which are used for making the decisions, and there is no chain of reasons that can easily explain how the algorithms came to these. The developers themselves can in most cases not explain how the algorithm made its decisions. Seeing what is happening "inside" the algorithm is almost impossible, leading to what is called the "black box problem" or "the opacity problem" of AI (Eschenbach, 2021; Danaher, 2016). Furthermore,

algorithms might be not only technically opaque but also theoretically opaque, as very little is still understood about why these systems work so well and how (Beckman et al., 2022). These factors make it difficult to accurately explain how an algorithm is reasoning and to judge whether the decisions to use an AI solution have been grounded in good reasons. The challenge is to satisfy the proceduralist requirement that algorithmic decisions and AI solution decisions be made public so that the reasons given for their decisions, use, and design can be assessed.

6.4.2. Outcome-based requirements of AI publicity

For outcome-approaches, it would be important to obtain the outcomes of algorithmic decisions to discern whether they were good. This includes the decision or predictions made and how they correspond to a particular performance metric for example. Furthermore, in the outcome-approach, AI solution legitimacy would necessitate the outcome of the solution according to the motivated measurement, alternatively would correspond to an evaluation of the outcome of the system on a larger scale. This would require transparency in the problem or policy that the AI was made to solve or implement and sufficient knowledge about its measured and unmeasured outcomes. The latter might be hard to come by.

6.4.3. Questions of accountability

The principle of publicity presents a way for citizens to hold the government accountable for their actions. To ensure the different kinds of AI legitimacy discussed above the proceduralist and outcome-approach requires citizens to have access to certain information and knowledge about the government's AI use. The developers and users of AI technologies are far from oblivious to the problems of algorithmic opacity and opacity in the decision-making process. Therefore, several solutions have been proposed for how to increase AI publicity and, ultimately, accountability of AI use.

6.5. Accountability methods

6.5.1. Explainable AI

6.5.1.1. Transparent and post-hoc methods

The need to provide explanations for the algorithm's decisions has given rise to the field of Explainable AI (XAI). XAI consists of different techniques for explaining algorithms' inner workings or reasoning. Therefore, it is relevant for the proceduralist account of legitimacy. XAI techniques are often divided into two categories: transparent methods and post-hoc methods. Transparent methods are algorithms considered interpretable by nature, meaning their inner workings can be truthfully represented and obtained for inspection (Gohel et al., 2021). Such algorithms include decision trees, Bayesian models, and linear regression. If the internal feature correlations are not too complex, interpretable methods are good solutions to problems, as they avoid the problem of opacity from the start.

In contrast, post-hoc methods are used to explain complex algorithms whose internal workings remain opaque. They receive a trained model as input and generate useful representations constructed from approximations of the inner workings and logic of the opaque model (Gohel et al., 2021). Depending on the data that the model to be explained uses (text, tables, images), the post-hoc methods generally try to explain what features in the data the model paid the most attention to when making its decisions. However, many of these methods are not robust and can therefore be misleading. An example is an XAI method that tries to explain what part of an image the algorithm paid attention to when classifying a picture as being of a dog or a cat. This technique might highlight the pixels the algorithm paid the most attention to when making its decision. However, as image classification algorithms use edges to determine the object in an image, the explanation provided by a post-hoc method might be the same regardless of whether the image was classified correctly or not. This means that XAI methods are not always reliable.

6.5.1.2. Challenges to XAI

Bruijn et al. (2022) investigated different uses of XAI methods in governments and identified several challenges to using XAI methods in the public sector. Firstly, citizens might not have the required expertise to understand the explanations provided by XAI methods. If an explanation involves how important different features were in the decision, it might still be difficult for an unknowledgeable citizen to understand if the decision procedure was fair. Therefore, an impartial body might be necessary to help with this judgment. Furthermore, work on combining XAI methods with conversational agents could be a potential (technical) solution to this problem (see Nguyen et al., 2022 for an example).

Another challenge with XAI is that algorithms change as they get retrained on new data. That means that the way that an algorithm reasons today might not be the same as last week. Suppose algorithmic harm occurred from a past model. In that case, the specific correlations that led to that result might be gone in the next update, which means that unsound reasoning that could have been caught by an XAI method does not remain if the past model and its predictions are not available. To assure accountability, it is therefore important that there is access to different versions of the models.

Another challenge for XAI is that the methods might be applied to legitimise potentially problematic uses of AI on difficult problems. By appearing responsible and transparent for its use of XAI, governments might get away with harmful uses of algorithms for difficult problems. Algorithms are often used to address so-called wicked problems in the public sector (Bruijn et al., 2022). These are problems that are difficult to solve because the relevant facts are ambiguous, and actors involved in solving the problem diverge in their opinions on whether a problem is important to solve and what would be a desirable outcome (Rittle & Webber, 1973). Examples could be hunger, income disparity, poverty, terrorism and sustainability. Solving wicked problems with AI solutions might imply only certain solutions which are easily quantifiable or for which data exists get considered. Even if there is XAI to explain such solutions, it does not explain the values that went into choosing that specific solution to the problem, which might be a requirement for proceduralists.

Because of these challenges, XAI methods might not lead to increased trust in the government's use of AI, but could sometimes do the opposite. This means that even if explainability of algorithms is useful for giving reasons according to the principle of publicity, the ideal of transparency might be limited if only considering the inner workings of the algorithms. The reasoning behind the application of an algorithm and the problem formulation also matter for AI publicity.

6.5.2. Human-in-the-loop

6.5.2.1. Human oversight and final decision

In order to ensure moral responsibility for algorithmic decisions in an institution, a common suggestion is the need to pair XAI solutions with a human-in-the-loop (Baum et al., 2022). The idea of a human-in-the-loop is that there would be a person overseeing the algorithm who can investigate and halt its decisions if perceived to be wrong. Alternatively, the human in the loop would be supplied with a recommendation for an action by the algorithm but would have the final say. By having human judgment alongside an algorithm, the idea is that it would be possible to exercise meaningful control and consequently ascribe responsibility for errors. An example is an algorithm used in the criminal justice system for evaluating whether a person should be released for parole based on their calculated risk of reoffending. The algorithm would be used to make the prediction, but the decision to give parole would ultimately be up to human judgement by a judge or parole officer (van Dijck, 2022).

6.5.2.2. Challenges to human-in-the-loop

However, Leins & Kaspersen (2021) argue that the ideal of a human as an overseer of a system is unrealistic as a practice. They argue that having any "meaningful interaction" with data or sensors is difficult, both at the time of data collection and operation. Because of this, it is impossible for a human to have oversight of all parts of such a complex system. Another problem with human-in-the-loop is that people generally get bored when performing a task that is not very challenging and does not require their full attention. This is particularly true for the supervision of autonomous systems. Furthermore, some understanding of the system is required, which would need to be paired with knowledge of the application domain so that it is possible to discern what are plausible reasons for a decision and whether the system was functioning correctly.

McQuillan (2022) writes that there are many reasons why people might defer to the decisions of an AI, even in situations where they doubt its decision. He thinks the most probable reason is that the human in the loop is aware of the institution's priorities. Suppose an AI solution is put there for efficiency reasons. In that case, questioning of a decision can be seen as "causing a problem" or show misalignment with the priorities of the institution. Moreover, the human in the loop has a heavy responsibility as they will absorb the blame for any malfunctioning. This puts the human in a difficult position where they must adhere to institutional priorities while carrying most of the moral and legal responsibilities when the system fails (Elish, 2019). This is especially troubling since even those with the correct skillset cannot always predict or prevent how the systems might malfunction or behave. Liens & Kaspersen (2021) further note that with human-in-the-loop, we

are still talking about a specific human, which might not be representative of everybody's interests, and thus might not stop decisions if they do not appear problematic to them.

The idea of human-in-the-loop might be desirable in order to assign responsibility. However, the actual effectiveness of human-in-the-loop for controlling unwanted outcomes might depend on the context of deployment, the knowledge and representativeness of the human in question, and the pressures from the internal organisation. From the proceduralist account of legitimacy, human-in-the-loop might not help increase legitimacy unless there is a possibility to discuss the decision to be made among people or if the human-in-the-loop is provably more fair in its decisions than the AI. From an outcome-approach, human-in-the-loop could help legitimacy if it increases good outcomes.

6.5.3. Transparency

6.5.3.1. Lack of transparency beyond XAI

XAI paired with a human-in-the-loop can be seen as a practical way to enable the principle of publicity for the public sector's use of algorithms. However, XAI methods only show the algorithm's inner workings or the relationships between variables and output. This is only transparency that might satisfy part of the proceduralist account of algorithmic decision legitimacy. It does so partly since the procedural legitimacy of algorithmic decisions might require an analysis also of the data. The opacity of algorithmic use in the public sector also stretches to input and output data, which is often unavailable for citizen inspection (Medaglia & Tangi, 2022). Furthermore, there is little knowledge on the part of citizens of where algorithms are applied, in what contexts, for what reasons, by whom and whether they are subjected to it. This means that it is difficult to evaluate AI solution legitimacy for both the proceduralist and outcome-approach. Beyond XAI, different levels of transparency would be needed to satisfy the publicity requirements of both accounts of legitimacy.

6.5.3.1. Challenges to transparency

Despite these extra levels of transparency, many have pointed out that publicity is not always in itself an unquestionable good. As Beckman et al. (2022) write: "unjust decisions are not more so just because they are public.". Ananny & Crawford (2018) have investigated several ways the ideal of transparency is limited and incomplete as a mechanism for algorithmic accountability. They argue that transparency can be harmful if it threatens privacy. Transparency can also intentionally be used to occlude if too much information is put out, making it hard to sift out the relevant information. Furthermore, they argue that transparency does not necessarily build trust and that it can prioritise seeing over understanding, which they argue is necessary if observers are to be able to debate and potentially change it. Transparency, then, is not an end in itself but a method for which its appropriateness must be evaluated in each context (Pozen, 2020).

Furthermore, the government may sometimes be unwilling to provide full transparency when it comes to algorithmic processes, as doing so also presents security or privacy risks. There are thus cases where transparency might enable the previously discussed adversarial attacks to the algorithms

or where vulnerable groups might be exposed. Another potential risk is that individuals with strong technical knowledge can use the information provided by full transparency to trick the system. However, one must ask how this differs from current systems, where a risk of individuals gambling the systems also already exists.

Hence, transparency for AI in the public sector might not be an end in itself but a means to an end. The question is what that end is. For proceduralist and outcome-based legitimacy, transparency is a means to hold the government accountable and allow the public to discern whether decision procedures or outcomes are legitimate. For the limits of the ideal of transparency, it might be important to consider what needs to be accessible and how it should be accessed for citizens to hold the government accountable.

6.6. Legitimacy conclusion

The problems with XAI, the limits of transparency, and the potential inefficiency of human-in-the-loop show the difficulty in establishing effective ways of upholding the principle of publicity for AI use in the public sector. This is problematic since AI technologies will be allocated decisions that the public might want to be able to judge if they are legitimate. The risk could be that such judgements are left to experts who understand the systems, which might mean that decisions and uses that the public would not agree with could be deemed legitimate. It is therefore important to involve the citizens in a discussion of how they would like to be governed by AI given these limitations and what solutions to publicity and accountability they would deem necessary. Furthermore, such a discussion would in itself contribute to AI literacy, which has been identified as an important missing piece in making the ideals of responsible AI work (Bruijn et al., 2022).

In this section, philosophical theories of democratic legitimacy were used to analyse whether algorithmic decisions, AI solution decisions, or an algocratic system could be considered legitimate. The proceduralist approach to legitimacy was challenged due to the opacity of the algorithms. For the outcome-approach, judging legitimacy would require specifying what outcomes to be considered and potentially giving reasons for why only these should be considered. Furthermore, the outcome-approach might lead to the legitimation of algocracy, which would be problematic from a proceduralist stance since it would minimise the opportunity for human participation. For a mixed approach, and with the problems documented for both the proceduralist and outcome-based stances, a topic for public deliberation could then be what the acceptable tradeoffs are between the two approaches and what is possible given the current limitations of the technology and the methods for AI publicity.

7. Debates in AI

I have argued that AI in the public sector impacts the democratic values of equality, justice, and freedom. This impact raised concerns about whether the use of AI in the public sector is legitimate. In the last chapter, I discussed different conceptions of legitimacy related to AI use in the public sector. I argued that, beyond publicity, citizen participation would be required to make the use of AI legitimate. Based on my findings, I conclude that AI in the public sector is a topic suitable for deliberation and one that should be deliberated to gain legitimacy.

In this chapter, I turn away from the investigation of why AI in the public sector should be deliberated to instead focus on how larger debates about AI as a technology might influence a public deliberation. This chapter gives an impression of how a public deliberation on AI could be influenced by two general debates in AI. Examining differing viewpoints and disagreements about AI can serve as a starting point for fruitful deliberation. Furthermore, these debates have significant ethical dimensions that interact with people's preferences regarding equality, justice, freedom and legitimacy. By including current debates on AI, I aim to contextualise the understanding of how citizens might reason about AI in the public sector.

7.1. Technological development and the question of control

A common debate about AI is whether its development can be controlled. This question can be discussed from the perspective of evolution, design, or both. I will discuss these different views and how they affect opinions about AI in the public sector.

7.1.1. AI development as evolution

7.1.1.1. An autonomous force

One common way of conceptualising technology is to think of it as an autonomous force - this theory is called *technological determinism* and has mainly been furthered by Jacques Ellul (Coeckelbergh, 2020). Ellul argued that whereas technology was once a means to human ends, technology has now become a reality in itself, supplanting the old reality of humans living with nature to one where humans are now subject to the organising force of technology as it progresses (Ellul, 1964). The internet can be seen as an example of this, as much of life has been organised around the internet, shaping the next phase of human development. Technological determinism claims that technology is the force that organises society in a way that suits its progression, which means that political, cultural, or economic forces can only exert control to the extent that it aligns with technological evolution. According to Ellul, technology is an all-absorbing, totalitarian force we cannot escape.

Technological determinism is present in current debates about AI. A doomsday understanding of the autonomous AI that escapes human control has been propagated in mainstream culture through the popularity of films such as *Terminator, Ex-Machina,* and *2001: A Space Odyssey* for example, but the same fear is also voiced by academics, such as Stephen Hawking's warning that AI

could end humanity (Higgins, 2018). Some see an autonomous AI as a force for good that will bring progress and prosperity to humankind and even allow us to transcend our bodies, as in the transhuman movement (*Mission*, n.d.).

7.1.1.2. Development towards AGI or superintelligence

In the current discourse on AI, it is often assumed that the trajectory of AI development is known: that it will inevitably develop towards higher and higher levels of intelligence. Many believe that AI can become as intelligent as humans and beyond (superintelligence) and are actively pursuing that goal. This is the project of creating "strong AI" (a system that exhibits human-level intelligence), which is different from the project of creating "weak AI" (systems designed to perform specific tasks) (Russel & Norvig, 2021). Whether strong AI is possible or not is a debate which stems from the question of what "hardware" is necessary to create intelligence. In this debate, some argue that "real" intelligence can only stem from a human brain, whereas others contend that intelligence is the same as computation, and can be created by machines.

Many are now arguing that AI is developing towards strong AI, also commonly called AGI - Artificial General Intelligence (Schwartz, 2018). Some people believe that once AGI is achieved, there will be a moment when there is a transformation from AGI to superintelligence: when machines will become more intelligent than humans. This event is often referred to as *the singularity*. It is hypothesised to lead to an explosion of ever-greater levels of intelligence, where each machine can create machines smarter than itself (Kurtzweil, 2005). This is assumed to lead to "enormous potential benefits": a cure for all known diseases, an end to poverty, and extraordinary scientific advances (Chalmers, 2010, p.4). Consequently, it is also said to lead to enormous potential dangers, with machines having the power to destroy the planet.

Since AI in this view has its own logic and agency, people could easily argue that any efforts to restrain AI will be pointless and that any developments of AI that are currently made and managed by humans will eventually be created and managed by the AI itself. Depending on whether one thinks that AI will be a force for good or for bad, the suggestions for how to manage AI in the public sector are likely to vary. If AI is thought to bring doom and existential risk, people might argue that the project to use AI in the public sector should be carefully analysed for potential detrimental outcomes and that safety guards should be put in place to diminish existential risks.

Suppose one instead believes that AI will develop towards a superintelligence that can manage society better than a human collective. In that case, people might argue that any efforts to control AI or try to understand it (for example through XAI) are futile since it is beyond human intelligence. This might lead to arguments that biases, even though a problem now, will be overcome with time as a superintelligent AI will eventually be able to eliminate bias through perfect data and hence render a truthful representation of reality (Moser et al., 2022). Biases might therefore be seen as temporary problems that can be solved with more data and better computational power. Because of this, they might argue that humans should try to hand over more and more tasks to AI and build the necessary infrastructure for it to evolve and eventually manage our societies. Hence, if AI is believed to be something unprecedented that will do more than solve

bounded problems, then any ethical and political discussion will likely become oriented towards a hypothetical future scenario (Cole et al., 2022).

7.1.2. AI development as design

7.1.2.1. Technology as a human product or social construct

Against the view that technology is an autonomous force that determines its own development, some might instead think of technology as a human product or construct. This view holds that human values and interests shape technologies, making them open to human choice through design (Poel, 2020). This also means that political, cultural, and economic forces can influence the direction of technology. Because this view acknowledges that values are embodied in technology, it also acknowledges that social powers can structure technological development and design so that if a dominant group has a certain interpretation of technology, technological change will take that path, leading to exclusionary design practices. An example is how bandaids often are designed to "blend in" with white skin in terms of colour and up until recently there were not many other colour choices for adhesive bandages (Wittkower, 2018). This normalises whiteness and defines other skin colours as deviants.

According to this approach, it is possible to control AI by designing it in ways that support human interests. This can be seen in Isaac Asimov's three laws of robotics, which try to prevent robots from hurting or disobeying humans while protecting themselves (Asimov, 1950). AI development through design is present in policy-making, for example in the AI Act which tries to control the use and development of certain AI technologies by classifying them into certain risk categories that are subject to different legal actions (Chini, 2023). Different AI systems can be designed to fulfil diverging understandings of equality, justice, and freedom. When discussing AI as a product of human design, people will have to argue for their view of what are acceptable choices for various issues like fairness metrics or surveillance. As this view holds that design can be used to control technology, there are as many views of what this could mean as people holding this view. This category might also include people who believe that AI will eventually become superintelligent, even though it is not now, and might acknowledge that humans could still try to exert some kind of control of its direction in the present. This opens up questions about how the risk of losing control in the future should influence how AI is developed in the present. Such discussions usually centre around the need to align AI with human values, which is often referred to as the *value alignment problem* (Gabriel, 2020). Public deliberation could help identify what values or principles a society could reach a consensus on and therefore endorse. Hence, even if people believe that AI will eventually become superintelligent, they might contend that there is room for influence at this point in its development and therefore welcome public deliberation about AI as a means to establish what values AI should (try to) be aligned with.

7.1.3. AI development as co-evolution

The third approach, called co-evolution, conceives of technological change as not fully autonomous but also not fully open to human design and construction (Poel, 2020). This approach argues that the second approach might overstate the degree to which technology can be

controlled by human interests. This is due to the non-malleability of technology, which makes the technology hard to govern and also means that technology creates novelty. It is not always easy to predict how a technology will unfold, and there will be unforeseen consequences and emergent properties of a technology. This is partly due to its technical nature but also because of societal factors like organisational complexities, economic considerations, social institutions and power constellations. This novelty means that it is not fully possible to capture a technology beforehand since it is difficult to predict if it will positively or negatively affect society. An example is how the release of the generative AI chatbot ChatGPT required the EU to revise their proposed AI Act regulation since it could not be neatly classified into a risk category (Volpicelli, 2023).

This approach states that technology and society co-evolve. The novelty and unexpected or unintended consequences in technological development give rise to the so-called dilemma of technological control (Poel, 2020). This dilemma describes that a new technology is still malleable in its early phases, but at that point, there is a lack of knowledge about its social impact meaning one does not know in which direction to steer it. Only later, when the technology has been more integrated in society, this knowledge becomes available, at which point it is difficult to change. This view is in line with technological materialism. AI can be understood in this way as it is sometimes hard to predict how it will affect society. This is due to its novelty and its reliance on optimisation in uncertain situations. This means AI sometimes creates unexpected effects that could not be foreseen, and therefore not controlled for. For example, algorithm biases that developers did not foresee could have resulted from representational bias in the data or the use of an inefficient proxy (Suresh & Guttag, 2019).

In a public deliberation on AI in the public sector, the co-evolutionary approach might argue for the need for more experimentation on AI solutions in order to be able to assess its potential outcomes before the technology is fully "locked in" to the system (Poel, 2020). Regarding fairness metrics, they might argue that in different domains, one could experiment with different metrics and try to assess their values and impacts. Social experimentation might also be another way that citizens could call for more AI solution legitimacy in the use of AI in the public sector since more than one solution would have been tested for. Since the co-evolutionary perspective recognises that society might change with technology, it is also more open to the moral change that technology might bring. This perspective can facilitate a discussion about what potential outcomes of different technological solutions (such as surveillance) could be desirable and on what grounds, i.e. the tradeoffs between competing values.

7.2. AI knowledge and the question of expertise

Another debate about AI is what kind of epistemic access AI has to reality and whether the knowledge produced by AI has epistemic value and why. This is often contrasted with human epistemic access and reasoning and the value of subjective experience and consciousness. The debate about what kind of knowledge AI creates compared to humans also leads to a debate about what this means for democracy and decision-making in general. Different opinions about the role of expertise or knowledge inform who or what people think should rule. In this section, I will look

at some different perspectives regarding human judgment and reasoning versus AI's knowledge production and discuss how different views on the merits of those might impact the deliberation on AI in the public sector. Many people might not adhere solely to one perspective but might see merit and value in both. This makes it even more important to conduct a public deliberation to see what people can and cannot agree on.

7.2.1. Knowledge through reason

7.2.1.1. Expert knowledge in the public sector

Since the end of the Cold War, governance has increased emphasis on using evidence-based methods for public policy analysis (Gilley, 2017). The complexity, risks, uncertainties, value pluralism, and expectations of public policy and the government have increased, leading to demands for "smart" public policies that can handle the newfound complexity efficiently and correctly. The idea was that, with data and means to process the data, one could determine how to direct policy and use data to measure its efficiency, only directing efforts towards worthwhile projects based on evidence. This was and is believed to lead to a better, more effective, and objective society, necessitating experts to discern how public policy could be realised through technological means (Broomfield & Reutter, 2022). This is a form of technocracy, which is rule by the skilled or the experts, in contrast with democracy, which is rule by the people. Technocracy can threaten democracy, as it might be difficult to distinguish between the outcome or aim of public policy and the means. Decisions that are usually made by democratic decision-making might come to be made by experts. Therefore, technocracy raises concerns about value conflicts and procedural legitimacy as it involves a lack of insight and participation in the process or a lack of justification for decisions made (Cole, 2022).

AI in the public sector can be seen as the next step in this call for evidence-based approaches as it uses mathematics, logic, and statistics to provide knowledge for decision-making, in many ways functioning as "expert" technologies. By being part of a larger system of people, organisations, computational networks, norms and practises, AI in the public sector contributes to the broader social endeavour of knowledge production - creating the knowledge necessary for taking political action (Yeung, 2022). This raises the question: what knowledge does it produce? Furthermore, is the kind of knowledge AI offers sufficient for decision-making?

7.2.1.2. Rationalism and algorithmic design principles

The difference between the perceived value of AI's knowledge and human knowledge has to do with two different ideas on how we gain knowledge of the world and is therefore mainly a question of epistemology. The epistemological theory that forms the basis for much of AI research today is rationalism (Shneiderman, 2021). Rationalism claims concepts and knowledge are acquired independently from sense experience (Markie & Folescu, 2023). Rationalists believe that the content of our concepts transcends sense experience and that these concepts can be grasped by reason. This supposes that reality has some structural properties that the intellect or the algorithms can grasp. Opposed to rationalism is the theory of empiricism, which argues that knowledge is gained mainly through sense experience.

Algorithms are designed according to rationalist principles in that they contain pre-programmed rules (code) that function as structures for creating knowledge. These structures can be used to make inferences that can be considered rational and sound. As rationalists believe in the power of logic, formalism, and abstraction to create clarity of knowledge and certainty, rationalists prefer algorithms to human decision-makers as algorithms use abstracted data about the world and create knowledge of this data through formal rules and mathematics. The reasoning of algorithms, as they concern only abstract representations about the world, are therefore thought to be more "true" than human reasoning and testimony.

Rationalists also favour the idea that statistical methods like machine learning are sufficient to create "strong AI" - AI that matches or exceeds human abilities (Shneiderman, 2021). They argue that if the data is rich enough, causal structures can be inferred from data. Hence, rationalists argue that more or better data is all that is needed to create a true or sufficient picture of reality. This solution is also proposed for current algorithmic errors (Chowdhury, 2023). The reliance on more or better data to create correct knowledge from reality sometimes also accompanies a claim that theories are no longer needed - the truth can be found in the data (evidence) with the help of algorithms (Mikhailenko & Dorrer, 2020). Therefore, the rational approach tends to reduce AI to a discipline of natural sciences, where empirical data is used to create formal models of the world that can be used to explain (and make decisions) about the studied phenomena (Ganascia, 2010).

When applied to politics, rationalism requires that policy questions be framed as rational decision processes for finding objectively better outcomes (Coeckelbergh, 2022b). This can be seen in the language of contemporary policy analysis where the norms of what defines "good" policy are that they should be "feasible; evaluable; benefit more than they cost; be effective in addressing some problem; be reasonably certain of success; be well grounded in evidence; and be amenable to monitoring and evaluation." (Gilley, 2017, p.13). As some argue that AI is a more rational decision-maker than humans, they might propose that AI should be used to identify these better outcomes. Sætra (2020) argues that one could defend a technocracy of AI on the premises: a) AI is superior to human decision-making in certain areas, b) This is particularly the case in solving complex problems, c) Politics is a complex problem; therefore AI can be used to "solve" politics. An example is an algorithm used to find tax policies that could effectively trade off economic equality and productivity in dynamic economies (Zheng et al., 2020).

7.2.1.3. Rationalistic views in public deliberation

The rationalist approach to AI development is present in many people who are currently considered "AI leaders" (O'Connell, 2017). As such, they have popularised rationalism as an approach to designing algorithms in particular and also as a "rational and empirical way of tackling every day" in general (McQuillan, 2022, p.93). Many people will therefore hold rationalist views in a public deliberation, emphasising that algorithms create better knowledge than humans by exhibiting reasoned decision-making that is deemed to be coherent and logically consistent. Rationalists might argue that more decisions should be ceased to AI on the basis of it creating more logically coherent knowledge than humans (Sætra, 2020). They might therefore endorse an AI-led technocracy. However, some rationalists might still hold that the public could decide on the ends of

public policy as long as AI is used to identify the means. This requires a rigorous specification of what a good "end" is regarding equality, justice, and freedom, which could perhaps be easier to specify with the help of public deliberation. If consensus could be reached on what equality or justice should look like for a specific application or situation, then from a rationalist point of view, what would be required would be a translation of those values into measurements. This could pertain to which fairness metrics should be used or lead to ideas about new metrics that could measure other concepts, such as freedom.

Rationalists are more likely to take an outcome-approach to legitimacy, emphasising the measurements of outcomes to ensure AI is creating good solutions. From a rationalist point of view, a government that involved less human emotions or empathy through the use of machines might be considered more legitimate, as these could be argued to "get in the way" in finding the optimal solution. Therefore, some rationalists who align with technological determinism might argue that public governance should be given to a superintelligent AI, which could manage without human input (Domingos, 2015; Hidalgo, 2018).

7.2.2. Knowledge through experience

7.2.2.1. Empiricism

In contrast to rationalism, empiricism argues that human knowledge is gained mainly through sense experience. This includes both sense experience, involving the five senses, and reflective experience, such as conscious awareness of one's mental operations (Markie & Folescu, 2023). For empiricists, sense experience allows one to acquire knowledge of external objects, and reflective experience helps establish this knowledge in the mind.

Empiricists might argue that some decisions should be left to humans in certain domains because of the expertise and context-sensitivity that comes from lived experience. Some might make a lesser claim and say that humans should at least be involved in decision-making as they have important knowledge to contribute from their subjective understanding. This is a kind of subjectivism, an empiricist theory stating that the only thing we can know is our subjective experience (Merriam-Webster, n.d.). Regarding ethics, subjectivism posits that moral judgments are subjective and based on individual preferences or cultural norms rather than objective truths. They might therefore argue that any outcome from AI needs to align with people's lived experiences and understandings of a situation. According to empiricists, knowledge created through scientific methods, such as AI knowledge, might be necessary for political judgement but is insufficient as politics also requires normative judgement based on sense experience.

7.2.2.2. The problem of induction

Whereas rationalists focus on finding truth or certainty by generalising from large amounts of data, empiricists are more interested in contradictions and ambiguities that come with real-world experiences (Shneiderman, 2021). They might therefore be interested in the anomalies or unexpected cases in data and work to give sense to observations of complex individual cases (Ganascia, 2010). This means that empiricists might be against any decisions taken purely by

inductive inference. Inductive inference is the use of a finite set of past data (observations) to generalise about the future, which is one of the most common techniques in AI for decision-making (Romeijn, 2011). An example is predictive policing, which uses data about individuals to analyse risk factors like previous arrests to predict who will commit a crime next (Lau, 2020). This means that patterns of the past are assumed to be true also in the future.

Inductive methods are subject to the problem of induction, which states that one cannot be justified in one's belief that things that have happened in the past will happen again in the future (Henderson, 2022). Within a political context, decision-making with the help of inductive AI knowledge assumes that society will be relatively stable. This means that the past gets used as a blueprint for the future, which can be problematic if society has changed (for knowledge production) and if society should change (normative consideration). Even though humans also make use of inductive reasoning, empiricists might question whether political decisions should only be based on the probable, as opposed to the possible (McQuillan, 2020). The possible arises from people's imaginations and through deliberation, hence through both subjective and communal experiences. The possible is the visions that people have of what world they want to live in. This is a kind of knowledge that is relevant for political decision-making that is not dependent on the past.

7.2.2.3. Data as subjective

The emphasis on the value of the human subjective experience for making sense of the world means that empiricists argue that there is a gap in algorithmic logics that requires human judgement and deliberation in order to fully do justice to the plurality of human lived experience. According to empiricists, this could not be filled by more data as data is not objective. Rather, data is created from observation, which includes the choice of object, a specific task, a specific interest, point of view, and problem (Mikhailenko & Dorrer, 2020). Consequently, there is always a theoretical perspective involved in data, and empiricists would argue that the lack of causal structures in AI technologies limits the interpretation of data and can therefore lead to questionable conclusions about the results. An example is a facial recognition system developed at Harrisburg University, which was claimed to predict whether someone would become a criminal by looking at their face (Fussell, 2020). These claims were based on unsound scientific premises that many studies had already debunked (Coalition for Critical Technology, 2021) but speak to the danger empiricists see in using data without theoretical understanding. Some AI researchers have therefore started to include causal inference in AI technologies to make algorithmic results better grounded in theoretical knowledge (Pearl et al., 2016).

7.2.2.4. Empiricism in public deliberation

In a public deliberation, empiricists will likely argue that AI technologies in the public sector must be paired with human judgement (Alon-Barkat & Busuioc, 2021). That could mean they would call for a human-in-the-loop, since algorithms might not recognise hidden biases or absences of expected patterns, which humans potentially could. Some empiricists might be against using AI technologies, arguing that politics should stay a human affair and that emotions and subjective experience have important roles to play in politics (Westen, 2008). A public deliberation could then help establish what specific part of politics they would require to be left to human oversight or final

say, and what tasks they could conceive of leaving to algorithms and on what grounds. Furthermore, if empiricists are suspicious about the justifications of algorithmic decisions, they would likely argue for stronger AI solution legitimacy, which could include an appeal to use third-party auditing or citizen involvement (Jankin et al., 2018). Empiricists would also likely argue that positive freedom as freedom of participation needs to be protected, which could mean that they would promote public deliberation to establish public opinion before any use of algorithms to collect and analyse data about citizens' preferences and opinions. Therefore, empiricists are more likely to prefer democracy over a technocracy, although some might be comfortable with a technocracy consisting of human experts.

7.2.3. A combined account of knowledge creation

Ganascia (2010) argues that many of the failures of AI are caused due to the outer environment in which they were placed. In essence, this means context. This is also a common criticism of the rationalistic design principles of AI: by reducing components into separate atomic entities, it removes a particular object from its context, thereby losing important information. Data is a digital representation of objects or relations in the world which are deemed meaningful. However, only if data actually captures what we believe is important can AI performance be relevant. This means that even if an algorithm has a 99% accuracy on its dataset, this has little value unless this dataset also represents reality in a way that is accurate or useful. This also includes ethical considerations of how people should be treated in technological processes. An example is if an algorithm for allocating public housing units used information about a person's social network to allocate public housing. Here, the need for public housing was not evaluated in terms of income level, family size, or specific needs, which might be relevant factors to determine eligibility and instead used information about people's networks, potentially only distributing public housing to those who did not have a sufficiently strong network to turn to for support.

Furthermore, AI does not only model a specific situation but also uses the model to create new conditions in the world. These effects do not happen in a vacuum but in the real world, filled with complexity and always in a context. Empirical understanding is also needed to evaluate how AI affects people's lived experiences. Therefore, the rational and empirical approach to understanding AI is necessary, making AI an interdisciplinary study. This is further why a public deliberation on AI in the public sector is necessary; so that different perspectives can come together and learn from each other. Only then are AI technologies in the public sector likely to create outcomes that are good according to rational criteria and good according to the citizens in their first-hand experiences of the technology.

From a deliberative democratic point of view, legitimacy is gained through deliberation, which is likely to lead to epistemically sound outcomes because diversity trumps ability theory. This means expertise should not necessarily have political authority. Estlund (2004) calls this the expert/boss fallacy, which means that even though some people have more expertise than others "it simply does not follow from their expertise that they have authority over us, or that they ought to" (p.3). However, even if AI is not conceived of as an expert technology, it is evident that it is still useful for many kinds of decision-making and public management. That leaves open the possibility that both human knowledge and knowledge gained by AI can be included in the democratic process. A

public deliberation would be a good first step in figuring out what that could look like and what it would require.

7.3. Debates in AI conclusion

Public deliberation on AI will likely be influenced by these two larger debates about AI and bring forth different ideas about how the development and the knowledge-creating activities of AI will influence society and politics. Possible outcomes of deliberation for these debates are as follows: Technological determinism together with a valuation of AI's rationalistic knowledge creation might suggest that AI in the public sector will contribute to the shaping of society in a way that means that human behaviour will be determined by a new form of knowledge creation that follows rationalistic principles. Using AI will create new knowledge, which could also make people aware of new political decision points that could contribute to a better society. The call for action in this view might be preparedness for the coming change. Technological determinism with an empiricist view might instead focus on the way that human perceptions and experiences might create different evaluations of the effect of AI, highlighting the need for different evaluation methods to assess what stances one could take about AI development.

In a view that understands technology as a human product or construct open for design, a rationalistic view might emphasise the need to design AI according to rationalist ethical principles. This may mean an emphasis on making AI perform moral reasoning along ethical frameworks to ensure it aligns with human moral values. Such a view might seek to design AI technologies in a way that ensures human agency and control while still relying on AI's reasoning to solve particular problems. The understanding of technology as a human construct paired with an empirical valuation of knowledge creation might instead focus on designing technology according to people's experiences, focusing on personalisation and user-centred design. This approach would highlight the need for feedback between user and technology and user testing to evaluate how the technology impacts human experiences, behaviours, and values.

Since AI is likely neither fully designable nor fully autonomous, another view would highlight the necessity of taking both rationalistic and empirical approaches to AI development in the public sector to steer it while still expecting unintended consequences. This view might argue for focusing on designing AI according to rationalistic principles, including reasoning along ethical frameworks, and setting up structures for evaluating and monitoring of AI systems to catch unintended consequences. The empirical understanding would highlight the need to assess how these systems and their unintended consequences shape society and would argue for the need to incorporate structures that still allow for human decision-making and technological control. One such empirical method of assessing AI's impact on society could be public deliberation. How a public deliberation about AI use in the public sector could work in practice will be discussed next.

8. Public deliberation in practice

The ideals of deliberative democracy are aspirational and difficult to achieve in practice. One of the main difficulties is the tension between two equally important tenets of democratic legitimacy: deliberation on the one hand and mass participation on the other (Landemore, 2022). A famous model of how this would take place is Habermas's two-track system. Track one in Habermas's model is characterised by public deliberation in formal political institutions like the Parliament, the Courts, and administrative agencies (Habermas, 1996). This is where structured deliberation would occur as an exchange of reasons among parties in order to justify laws. The second track is instead characterised by a lack of structure where deliberation takes place in the larger public sphere among different groups and networks. This is a kind of "deliberation in the wild". Habermas's idea was that the discussions taking place in track two should set the agenda for the deliberation in track one, meaning that the larger public debates would inform policy decision-making.

The two-track system has been an influential theory for modern democracies. However, it does not necessarily fulfil the ideals of deliberative democracy since it separates those in power and the people. This means that there is no direct way to ensure that those in power are truly representative of the people or can give reasons that the people would accept. That raises the question of how one could better structure public deliberation to ensure quality deliberation and mass participation and representation.

One common suggestion to ensure both quality deliberation and representation is to use random sampling of the population to form so-called "mini-publics" that work as a representative sample of how the population at large would think about an issue (Landemore, 2022; Fishkin, 2018). The mini-publics are educated through briefing materials and plenary sessions with experts about the issue and asked to deliberate. I will argue for the use of assemblies of mini-publics as the best way to discern what the public thinks about AI use and development in the public sector. Crucially though, there needs to be a greater debate about AI going on concurrently with the deliberative assemblies where the whole population is invited to participate. Inspiration for this larger debate can be taken from (and learned from the shortcomings of) The Great National Debate that took place in France over two months in 2019 (Collier, 2019). The Great National Debate included grievance books in town halls where citizens could write their complaints or wishes, town hall meetings on different topics, and an online consultative platform where citizens could upload their proposals and engage in discussion. There were also 21 randomly selected regional assemblies of 100 participants who were asked to deliberate on the topics (Landemore, 2022). Except for the regional assemblies, the other methods were not necessarily deliberative. Rather than being examples of "an exchange of reasons among equals", they are more like the "deliberation in the wild" as envisioned by Habermas. Moreover, except for the regional assemblies, all of the methods relied on self-selection, meaning that it was mostly a specific demographic that turned up to the town meetings, meaning the sample was not representative.

A public deliberation on AI should aim to give everyone the opportunity to participate, as this is in line with the larger mission of increasing the AI literacy of the population. Therefore, public meetings, online deliberative platforms, media debates, and other communicative channels need to

be put in place. However, deliberative democracy also aims to give legitimacy to the decisions taken, which means there needs to be quality deliberation among equals, which can only happen when there is representation and good conditions for deliberation. I argue that mini-publics present the best option to fulfil these criteria in practice. In the following sections, I will present some design considerations for how those mini-publics could be structured and discuss some challenges to public deliberation in general. Importantly, the structure I present applies to a deliberation that aims to provide general guidelines for AI in the public sector and does not pertain to possible deliberations about specific AI projects.

8.1. Design deliberative mini-publics

8.1.1. Demographic and attitudinal representativeness

Representation is a crucial aspect in a mini-public since it is assumed that the population would come to the same conclusion if given the same chances to deliberate under good conditions (Fishkin, 2018). Representation concerns both demographic representation and attitudinal representation. Demographic representation includes many of the standard demographic categories, such as age, gender, ethnicity, education, and income. Attitudinal representativeness is the representation of different views or attitudes towards an issue and is not always considered important to account for in deliberation. In the case of AI in the public sector, one could argue that attitudinal representativeness is important, especially since deliberation can create innovative proposals, but only if there is a disagreement between participants (Ayano, 2021). Attitudinal representativeness is often evaluated with the help of a questionnaire to potential participants beforehand to ensure an even selection of different attitudes (Fishkin, 2018). Such a questionnaire should query participants' different attitudes towards AI and their attitudes towards the government. One could also argue that questionnaires could be used to ensure that a multitude of professional backgrounds is represented in a deliberation, ensuring that some people will have a more in-depth technical understanding to contribute with. Ensuring that the microcosm starts with similar viewpoints like the population contributes to the plausibility of the hypothetical inference that the reasons reached by the microcosm would also hold for the population.

8.1.2. Sample size

Representativeness also concerns sample size. A question is how large the total number of people to be distributed into mini-publics needs to be for any changes of opinions in the deliberation to be meaningfully evaluated (Fishkin, 2018). Having a large enough sample is important to guarantee that any changes are not just random noise. If the sample is not large enough to certify representativeness, there is less reason for the political decision-makers to consider the results legitimate and therefore act on them. Lafont (2015) argues that only mass deliberation would give political decisions democratic legitimacy. This is impossible in practice, and even for referendums, results are deemed legitimate if half the population participates (Landemore, 2022).

For public deliberation in mini-publics, it would be difficult to get 50% of the population to participate, as deliberation requires more time of the participants than one day of voting does.

Because of this, Landemore (2022) has argued that one could weaken this requirement to include 10-15% of the population to still get quality deliberation. This would still constitute mass deliberation and likely yield a representative sample size for most countries she claims. She argues that this sample would be legitimate if the mini-publics are also rotated a number of times, so that more of the participants talk to each other. Some might argue that if the sampling is well done, a much smaller proportion could still yield a representative sample that ensures quality deliberation. However, that might not be considered mass deliberation, meaning this ideal might have to be abandoned.

8.1.3. Incentives for participation

The recruitment of participants is a significant task to ensure representativeness and will likely need to include incentives. These could be both monetary and non-monetary. For non-monetary incentives, research by Jacquet (2018) on incentives that attract citizens to participate included a desire for sociability, the opportunity to learn more about the topic, and a sense of civic duty. Citizens were also motivated by the opportunity to influence the decision-makers and provide them with information about their concerns, thus functioning as a link between voters and politicians. For monetary incentives, Landemore (2020) proposes that an honorarium could be paid to participants to reduce the self-selection of asked participants. Another suggestion is for the organisers to pay for travel and lodging costs for the duration of the deliberation events (Gastil, 2017). This would be important in ensuring that those who cannot participate because of a lack of resources would still be able to participate in the deliberation, hence promoting the representativeness of the results.

8.1.4. Sufficient information and expert input

Given that the representativeness design criteria are satisfied, the merit of the conclusions of deliberation also depends on whether participants have had the opportunity to deliberate under good conditions. One such condition is sufficient information and knowledge about a question so that the participants can accurately weigh different reasons and arguments for or against each other (Fishkin, 2018). One way to ensure this is to put together briefing materials that give the participants an understanding of what AI is, its main capabilities, its limitations, and the challenges of AI to the topic of use in the public sector (risks to privacy, accountability etc.), and imagined opportunities and benefits of AI use in the public sector. This briefing material needs to be balanced, meaning it should address both the opportunities and challenges of AI. It should also be clear about what information about AI can be considered fact and what concepts are contested (for example the extent to which we can direct AI, or AI consciousness). Ideally, the discussion in small groups should also be alternated with plenary sessions where the groups can ask questions to experts. This facilitates knowledge gain during deliberations and helps participants reach an informed judgement.

AI has several debates about its progression, epistemic access, and usability. For this reason, it is important to involve competing experts. Beyond competing understandings about AI's progression or epistemic access, it would also be important to aim for demographic diversity of the experts. AI researchers and industry professionals do not represent the larger population as they are majority

cis-male, white, straight, and from affluent backgrounds ("Artificial Intelligence Index Report 2023," 2023; Crowell, 2023). Their lack of diversity regarding race and ethnicity, gender and sexual orientation also impacts their understanding of AI and its desirable use. By promoting a diverse set of AI experts in demographics and attitudes, the participants will likely get a more nuanced understanding of AI. It would also be important to involve not only technical experts, such as computer scientists, but also experts from sociology, philosophy, political science, law and public administration, for example. All these fields are already involved in interdisciplinary research on AI, and their insights are necessary to guarantee that participants' understanding of AI is as broad as possible.

Having competing and diverse experts also allows participants to consider competing arguments themselves, which can prevent them from simply deferring to expert knowledge (Fishkin, 2018). Deferral to experts is a common concern in deliberation since it would undermine mini-publics' function as a bridge between lay persons' judgements and experts'. However, there is some support that this worry is not warranted. For example, a study done by Leino et al. (2022) on a mini-public deliberation on Covid-19 measures in Finland found that even though expert opinion framed the deliberative process, it did not influence it in ways that meant systematic changes in opinion.

8.1.5. Mutual respect and deliberation capabilities

Another condition for a successful deliberation is ensuring that participants are able to express and justify their views in a respectful and equal way. This idea of mutual respect is central to deliberative democracy (Bächtiger et al., 2018). However, the ideal of mutual respect can be difficult to achieve in practice. There is a risk that deliberations are dominated by people with more privilege, people who are good at articulating their thoughts and opinions, or people who hold very strong views. This means that the inequalities that exist in the world get imported into the deliberation process, undercutting attempts to deliberate based on mutual respect.

Participating in deliberation requires some capabilities that are not distributed equally in the population. Sorial (2022) recognises three distinct capabilities required for deliberation. The first is the ability to formulate authentic preferences. If there are asymmetries in power and resources in the deliberation, participants might assent to the opinions of those more powerful or articulate, adapting their preferences without carefully considering whether they agree with them, therefore not forming their own opinions. The second capability is the effective use of cultural resources, meaning the ability to adapt speaking style to the culturally dominant way of expressing oneself in the deliberation context. Deliberation requires reason-giving and argumentation that ideally should lead to consensus based on *the unforced force of the better argument* (Habermas, 1996). However, what are considered good ways of making arguments is often culturally determined (Young, 1996). This means that people who cannot express their ideas in this way can get their arguments discredited or their voices silenced, undermining the representativeness of the outcomes. Thirdly, deliberation requires basic cognitive abilities and skills to articulate and defend persuasive claims (Sorial, 2022). Those that are well-resourced in terms of education and high incomes tend to participate more in formal politics and are therefore more skilled, leading those lacking these to self-exclude or assent rather than contribute with their viewpoints.

These inequalities can lead to “hermeneutical domination”, which is when participants from minority groups’ testimonies are dismissed because the majority groups perceive the group as a whole as epistemically untrustworthy (Catala, 2015). To counter this phenomenon and the potential for advantaged individuals to dominate the conversation, Catala suggests that deliberation needs strict rules that encourage participants to listen carefully, speak respectfully, be responsive to others’ contributions, and be self-critical. Facilitators or moderators can also be used to ensure these good conditions of deliberation. Facilitators are usually trained not to give any hint of their own opinions. They help the discussion along the set agenda by ensuring everyone gets heard and communicates respectfully (Fishkin, 2018). However, facilitators differ in their involvement which can influence the deliberation in both positive and negative ways, meaning the role of facilitation is important to ensure deliberative quality (Dillard, 2013).

Beyond human facilitators, some have suggested that AI technologies can be used to ensure good conditions. An example is an online AI-assisted deliberation platform developed by Siu (2022) and her team, which had features like speaking queue, timed agenda, real-time transcripts, tracking of offensive language, and nudging to join the discussion or consider arguments. Software like that could potentially be designed in order to work in physical meetings as well. AI could further be used for speech-to-text transcription combined with machine translation. This could enable people with a different native language than the dominant one to participate in their native language, potentially increasing the speakers' intelligibility and hence deliberative capacities. This further speaks to AI’s potential to facilitate citizen participation, which is another way AI could be used in the public sector.

8.2. Agenda for public deliberation

The objective of many public deliberations is to gain legitimacy for new policies by involving the public directly in the decision-making. This often involves discussing the pros and cons of a suggested new policy or the best alternative out of three to four suggested policies (Gastil, 2017). For AI in the public sector and depending on the country, there might not be a new policy about AI use that is up for debate. Instead, the deliberation can be framed as the need to build legitimacy around the government's current and future AI use. For EU countries, the new (as of June 2023) legislation around AI could necessitate new policies or clearer guidelines around AI use in the public sector, which could be framed as policy issues to be deliberated.

The overall objective of a public deliberation should be to discern *how* the public would like to be governed by AI. This subsequently requires the public to understand how they are currently governed with and without AI, what are the benefits and risks of governance by AI (which includes an understanding of the current capabilities and limits of the technology), what the future suggested uses of AI in the public sector are (what is the vision for the technology) and how the public can hold the public sector accountable for their use of AI. The agenda needs to be tailored according to the laws and practices of each country.

The areas that should be touched upon in public deliberation on the public sector's AI use are summarised in Table 1:

Topic	Questions
Current practices with and without AI	• What are good (and bad) current practices or uses of AI in the public sector? • What situations warrant human judgement alongside AI judgement? • Are there any situations that are not deemed appropriate for AI solutions?
AI treatment and equality	• How should AI treat individuals and groups (in general and specific contexts)? • How should AI handle current inequalities?
AI imposed restrictions	• What should AI enable the citizens to do? • What should AI not interfere with? • What kinds of interferences would be acceptable, and on what grounds?
Legitimacy and accountability	• What conditions make AI decisions and decisions to use AI legitimate? • What are the best options for holding the public sector accountable for their AI use? • How should the citizens participate in the development of AI in the public sector?
Vision building	• How could AI serve the citizens' needs? • What future potential technological uses would be worth exploring? • What is the best and worst-case scenario for AI use in the public sector (both short-term and long-term)?

Table 1. Agenda for a public deliberation on AI use in the public sector

The deliberation in mini-publics will be specialised towards AI use in the public sector and should therefore aim to be as practically informed as possible. However, there are also larger debates about AI that do not necessarily need to be structured to create specific results. A national debate on AI should include a discussion on the public sector's use of AI but can also include topics on AI and its impact on society in general. These discussions could be more open-ended and philosophical, touching on issues such as robot consciousness and treatment or future imagined development. AI can also be discussed in relation to different sectors or topics, such as AI's impact on the job market and suggested paths forwards, or how AI can be used for education. These debates are important for helping citizens create different visions about how AI should fit into their lives and society.

8.3. Outcomes

The outcome of the deliberative mini-publics should be a report on the aggregated results of the procedure. This report should include the prominent arguments for and against different AI uses (and their surrounding infrastructure) and clarify where consensus was reached. It should also show what issues were recognised and need further discussion. If the citizens can make concrete suggestions that could translate into policy, these should be framed as preliminary policy suggestions. This report should be presented as a decision basis for the policy-makers in their AI use and plans. If the outcomes can be summarised as policy suggestions, these could serve as a basis for a national referendum. The onus is then on the policy-makers to adhere to the citizens' decisions.

The most important outcome of a greater national debate about AI is that it would familiarise and educate the public about AI. Such an event could therefore help organise the collective to gain more political power. Dan McQuillan (2022) has proposed people's councils as a means to give political traction to the talk of "centring the voices of the marginalised" that is often discussed in AI without being further developed. He imagines this council as self-constituting, without being granted authority by any institution, and therefore not assuming loyalty to any public body. Public councils could therefore be important as a means to safe-guard and monitor any potential technocracy created with AI.

For the private sector, public councils could be supplemented by worker's councils on AI, where tech workers organise to gain more decision-power over what technologies they help develop. The worker's councils could then work as a way of holding the private sector accountable from the inside. A greater national debate on AI could also help the private sector become more accountable to their customers by creating products that fit their needs and create accountability measures that are accessible to the customers.

8.4. Challenges to public deliberation

8.4.1. Manipulation

One threat to public deliberation on AI in government is manipulation, and one potential source of manipulation is the government itself. The government could try to steer the deliberation by framing the questions to invite a specific perspective on AI while omitting others. The government could also be explicitly involved in selecting experts, which could mean a selection of experts that are particularly aligned with the government's vision of AI. However, manipulation of a public deliberation is also likely to come from the private sector, individuals or groups in society, or international actors. Many stakeholders have strong incentives to steer opinion in a specific way, as the public sector has abilities to make their decisions into policies.

If it cannot be accurately assumed that participants have control over their own political beliefs, the legitimacy of a deliberative procedure could be undermined. Large-scale influencing with the help of targeted advertisement has been seen before in politics. An example is how Cambridge Analytica mined thousands of people's Facebook accounts without consent to create targeted ads for Donald

Trump's election campaign (Rosenberg et al., 2019). The access to large amounts of data means that those who develop, use, and own AI have a lot of control over people and are able to influence whatever is seen as the "truth" about AI. This is problematic since political agency relies on epistemic agency, which is control over one's beliefs and how these beliefs are formed and updated (Coeckelbergh, 2022a). If one's epistemic agency is threatened, people might also argue that any opinion a participant holds is not truly representative of the population at large. This brings one back towards a technocratic vision, where the public is seen as too uneducated to meaningfully participate in the decision-making.

Despite the risk of manipulation, the effects of manipulation on a deliberative mini-public might be mitigated by the design of the procedure itself. Deliberation not only relies on opinion exchange but also forces the participants to give reasons and justifications for their views. Deliberation has regularly been shown to lead to opinion change, and polarising effects are commonly eliminated by focusing on good deliberation design (Fishkin, 2018; Gastil, 2017). Use of trained facilitators, exposure to outside scrutiny, rotating the groups, and giving anonymous feedback are all methods that could increase the epistemic agency of participants and dispel outside manipulation.

8.4.2. Lack of follow-through

One challenge to public deliberation is for the results of the deliberation to be considered and influence exerted in the policy-making process (Dryzek & Niemeyer, 2011). Even though there are many potential roles that mini-publics could fill (for example directly making policy), for most deliberative bodies that convene on scientific issues, the outcomes are strictly advisory (Gastil, 2017). Even though there are many cases of mini-publics successfully exerting influence, there are many more examples where the decisions from the deliberations had no or little influence on public decision-making (Dryzek & Niemeyer, 2011). The roles that mini-publics play also vary greatly depending on political system.

Moreover, there is evidence of a systemic difference between what a deliberating public would conclude and what public decisions are reached regarding the risks associated with new technologies (Dryzek & Niemeyer, 2011). This evidence comes from comparative studies on mini-publics convened about the issue of GM foods, but there is reason to suspect that a similar effect could be seen for AI. The reason for this is that the public tends to be more precautionary than the policy-makers, meaning the burden of proof for the safety for the new technology is put on the proponents of the technology (O'Riordan & Cameron, 2013). In contrast to regular citizens, the policy-makers are often motivated or subordinated to the drive for economic growth and efficiency. This is especially true in times of austerity, which means that risks that are considered too high in the eyes of the public are considered acceptable to the policy-making elite as the risks are expected to be managed by the social and economic systems in due time (Dryzek & Niemeyer, 2011). As some of the most cited justifications for AI in the public sector is efficiency, it might be challenging for any deliberative outcome to exert actual influence on public decision-making processes if their suggestions are cautionary.

The lack of influence is problematic as public deliberation is seen as a method for building social accountability by promoting citizens' voices. Empirical evidence shows that information alone does

not necessarily motivate collective action or the government to act, which points to the need to go beyond promoting citizens' voices (Fox, 2015). For social accountability to work, there also needs to be strategies in place to respond to the citizens' voices. Without "teeth" to the voice, social accountability remains incomplete. As AI risks impacting people's lives in ways that disrespect the democratic values of equality, justice, and freedom, these "teeth" are owed to the citizens, and can be demanded through collective political action.

9. Conclusion

By showing the political nature of AI, this thesis has investigated how AI in the public sector might impact democratic values and argued that its potential negative impact raises questions of legitimacy that can be addressed by public deliberation. Even though often analysed purely in technical terms along an instrumentalist understanding, AI is not just a technical artefact but also comprises social practices with real-world impacts. This broader understanding of AI as a system raises questions about how AI influence democratic values when used in public governance.

Algorithmic decision-making can introduce biases due to historical bias in data, inadequate data collection procedures, or incomplete problem formulation and understanding. This can lead to citizens being treated differently due to their group membership, potentially leading to algorithmic harm and discrimination at scale. This means AI risks disrespecting the democratic values of equality and justice. Taking a neutral stance to these biases by building more accurate tools is impossible, as biases from the past get propagated into the future, reinforcing the status quo. Furthermore, even if biases are addressed through technical solutions like fairness metrics, AI might still create substantial inequality, raising the question of how the disproportionate impact of AI can be mitigated to ensure that marginalised people do not bear all the burdens. Furthermore, algorithmic discrimination can also lead to people not receiving the resources they are entitled to from the government, imposing on their ability to control their life and hence impacting freedom. AI surveillance and the potential for algorithmic nudging also raise concerns about privacy and paternalism, leading to questions of how one can balance the democratic value of freedom against other values like public safety and order. The risk of misuse of AI technologies and the lack of insight and citizen participation in its development process raises concerns as to whether the uses of AI and their decisions can be considered legitimate. Different conceptions of legitimacy have different conceptions of what would make AI use in the public sector legitimate. Whereas it is possible to evaluate legitimacy based on the outcomes AI creates, this risks creating a technocracy where only experts can evaluate the outcome of the algorithms based on a narrow analysis of how well they represent reality according to their data. I argue that this analysis is too narrow and that public deliberation is the best way to obtain legitimacy for AI use in the public sector. Public deliberation on AI use in the public sector should investigate how the citizens would like to be governed by AI and how different values can be weighed against each other. For a good outcome of public deliberation, I have suggested the form of mini-publics, which can ensure good conditions for deliberation while remaining representative of the population. A difficult challenge for public deliberation would be ensuring that the results can influence the decision-making on AI use in the public sector.

The political nature of AI can be seen in how it interacts with society and reshapes social relations, especially between citizens and the state. It is not sufficient for the government to provide details about where they use AI and why. The wide-ranging implications of these technologies on the democratic values of equality, justice, and freedom mean that citizens risk being negatively affected in ways they are not aware of and have not consented to. As such, the citizens must be involved in steering AI in the public sector, ensuring that its use not only contributes to making the public sector more efficient but also to values and efforts that the citizens collectively decide on. Public

deliberation on how AI can and should be used in the public sector is a good way to democratically decide on AI's direction. Involving citizens in deciding how AI should be used in the government can help us create a future society that is not only limited by the probable but instead aims for what we collectively imagine as possible.

Acknowledgements

I would like to express my heartfelt gratitude towards my project supervisor Dr. Dominik Klein who was patient with me throughout the many twist and turns of the project. Dominik provided me with great feedback and with interesting and fun discussions that have deepened my understanding of the philosophy of AI and political philosophy. I am also grateful to my second examiner, Dr. Johannes Korbmacher, for taking the time to read and assess this thesis. I am also grateful to Utrecht University and the Utrecht University Fund for providing me with the Bright Minds Fellowship, enabling me to obtain my Master's degree at a world-class university, and inspiring me to use my knowledge to continue to do good in the field of Artificial Intelligence. Finally, I am grateful to all the people who supported me throughout this process by listening to me complain, aiding my understanding by discussing concepts, giving feedback and proof-reading my work, and also for simply feeding me in the final stretch. I want to thank my parents, Cagin, Marta, Helin, Niki, and Nilüfer for this support. Also, I want to give a final shoutout to myself, for always making life a little bit harder, but also a little bit more interesting.

References

Ackerman, M.S. (2000). The Intellectual Challenge of CSCW: The Gap Between Social Requirements and Technical Feasibility. Human-Computer Interaction, 15, pp. 179 - 203.

Ali, A.T., Abdullah, H.S., & Fadhil, M.N. (2021). Voice recognition system using machine learning techniques. *Materials Today: Proceedings*.

Alon-Barkat, S., & Busuioc, M. (2021). Human-AI Interactions in Public Sector Decision-Making:"Automation Bias"and"Selective Adherence"to Algorithmic Advice. *Journal of Public Administration Research and Theory*, *33 1*, pp. 153–169. https://doi-org.proxy.library.uu.nl/10.1093/jopart/muac007

Ananny, M., & Crawford, K. (2018). Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. *New Media & Society, 20*, pp. 973–989.

Ananny, M. (2021). Presence of Absence: Exploring the Democratic Significance of Silence. In Bernholz, L., Landemore, H., & Reich, R. (Eds.), *Digital Technology and Democratic Theory*. University of Chicago Press, pp. 141-166

Anderson, E. (2012). Equality. In D. Estlund (Ed.), *The Oxford Handbook of Political Philosophy*, Oxford Handbooks. https://doi-org.proxy.library.uu.nl/10.1093/oxfordhb/9780195376692.013.0002

Anderson, J., Bholat, D., Gharbawi, M., & Thew, O. (2021, February 24). *The impact of COVID-19 on artificial intelligence in banking*. Bruegel | the Brussels-based Economic

Think Tank. Retrieved June 6, 2023, from https://www.bruegel.org/blog-post/impact-covid-19-artificial-intelligence-banking

Araya, D. (2015, September 14). Interview: Tim O'Reilly talks algorithmic regulation, cyber-terrorism, and why he doesn't like the term "automation." *Futurism*. https://futurism.com/interview-tim-oreilly-talks-algorithmic-regulation-cyber-terrorism-and-why-he-doesnt-like-the-term-automation

Arneson, R. (2013). Egalitarianism. In E. Zalta (Ed,), *The Stanford Encyclopedia of Philosophy* (Summer 2013 ed.). Stanford University. https://plato.stanford.edu/archives/sum2013/entries/egalitarianism

Artificial Intelligence Act. Proposal 2021/206. European Parliament, Council of the European Union. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A52021PC0206

Artificial Intelligence Index Report 2023. (2023). In *Artificial Intelligence Index*. Retrieved June 28, 2023, from https://aiindex.stanford.edu/

Asimov, I. (1950). *I, Robot*. Roc.

Ayano, T. (2021). A survey of methods for evaluating mini-publics. *Asia-Pacific Journal of Regional Science, 5*, pp. 1–19.

Barocas, S., Crawford, K., Shapiro, A., & Wallach, H. (2017). The Problem With Bias: Allocative Versus Representational Harms in Machine Learning. *SIGCIS Conference*. http://meetings.sigcis.org/uploads/6/3/6/8/6368912/program.pdf

Baum, K., Mantel, S.P., Schmidt, E., & Speith, T. (2022). From Responsibility to Reason-Giving Explainable Artificial Intelligence. *Philosophy & Technology, 35*, pp. 1-30.

Beckman, L., Hultin Rosenberg, J., & Jebari, K. (2022). Artificial intelligence and democratic legitimacy. The problem of publicity in public authority. *AI & SOCIETY*.

Berlin, I. (1959). *Two Concepts of Liberty: An Inaugural Lecture Delivered Before the University of Oxford on 31 October 1958.* Oxford : Clarendon.

Bertram, C. (2023). Jean Jacques Rousseau. In E. Zalta & U. Nodelman (Eds.), *The Stanford Encyclopedia of Philosophy* (Summer 2023 ed.). Stanford University. https://plato.stanford.edu/archives/sum2023/entries/rousseau

Bhuiyan, J. (2023, April 27). First man wrongfully arrested because of facial recognition testifies as California weighs new bills. *The Guardian*. https://www.theguardian.com/us-news/2023/apr/27/california-police-facial-recognition-software

Binns, R. (2017). Fairness in Machine Learning: Lessons from Political Philosophy. *Decision-Making in Computational Design & Technology eJournal*.

Blacksher, E., Diebel, A., Forest, P., Goold, S.D., & Abelson, J. (2012). What is public deliberation? *The Hastings Center report, 42 2*, pp. 14-7.

Borgmann, A. (1992). *Crossing the Postmodern Divide*. University of Chicago Press.

Bowring, F. (2015). Negative and Positive Freedom: Lessons from, and to, Sociology. *Sociology, 49*, pp. 156–171.

Broomfield, H., & Reutter, L. (2022). In search of the citizen in the datafication of public administration. *Big Data & Society, 9*.

Bruijn, H.D., Warnier, M., & Janssen, M. (2022). The perils and pitfalls of explainable AI: Strategies for explaining algorithmic decision-making. *Gov. Inf. Q., 39*, 101666.

Buolamwini, J. (2017). *Gender shades : intersectional phenotypic and demographic evaluation of face datasets and gender classifiers.* https://dspace.mit.edu/handle/1721.1/114068

Bächtiger, A. et al. (2018). Deliberative Democracy: An Introduction. In A. Bächtiger et al. (Eds.), *The Oxford Handbook of Deliberative Democracy*, Oxford Handbooks. https://doi-org.proxy.library.uu.nl/10.1093/oxfordhb/9780198747369.013.50

Carpini, M.X., Cook, F.L., & Jacobs, L.R. (2004). Public deliberation, discursive participation, and citizen engagement: A review of the empirical literature. *Annual Review of Political Science, 7*, pp. 315–344.

Carter, I. (2022). Positive and Negative Liberty". In E. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Spring 2022 ed.). Stanford University. https://plato.stanford.edu/archives/spr2022/entries/liberty-positive-negative

Castelnovo, A., Crupi, R., Greco, G., Regoli, D., Penco, I. G., & Cosentini, A. C. (2022). A clarification of the nuances in the fairness metrics landscape. *Scientific Reports*, 12(1). https://doi.org/10.1038/s41598-022-07939-1

Catala, A. (2015). Democracy, Trust, and Epistemic Justice. *The Monist*, 98(4), pp. 424–440. https://doi.org/10.1093/monist/onv022

Centre for Epidemiology and Evidence. (2022). *Rapid surveillance using PHREDSS - Epidemiology and evidence*. Retrieved June 21, 2023, from https://www.health.nsw.gov.au/epidemiology/Pages/rapid-surveillance-using-PHREDSS.aspx

Chalmers, D. (2010). The Singularity: a Philosophical Analysis. *Journal of Consciousness Studies*.

Chini, M. (2023, June 13). *"Protecting citizens against the risks": EU law on artificial intelligence underway.* The Brussels Times. Retrieved June 24, 2023, from https://www.brusselstimes.com/550889/protecting-citizens-against-the-risks-eu-law-on-artificial-intelligence-in-the-making

Chowdhury, H. (2023, February 7). Sam Altman has one big problem to solve before ChatGPT can generate big cash — making it "woke." *Business Insider*. https://www.businessinsider.com/sam-altmans-chatgpt-has-a-bias-problem-that-could-get-it-canceled-2023-2?international=true&r=US&IR=T

Christiano, T. & Sameer, B. (2022). Democracy. In E. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Spring 2022 ed.). Stanford University. https://plato.stanford.edu/archives/spr2022/entries/democracy

Coalition for Critical Technology. (2021, September 21). *Abolish the #TechToPrisonPipeline*. Medium. https://medium.com/@CoalitionForCriticalTechnology/abolish-the-techtoprisonpipeline-9b5b14366b16

Coeckelbergh, M. (2020). *Introduction to Philosophy of Technology*. Oxford University Press.

Coeckelbergh, M. (2021). Green Leviathan or the Poetics of Political Liberty. In *Routledge eBooks*. https://doi.org/10.4324/9781003159490

Coeckelbergh, M. (2022a). Democracy, epistemic agency, and AI: political epistemology in times of artificial intelligence. *Ai and Ethics*, 1 - 10.

Coeckelbergh, M. (2022b). *The Political Philosophy of AI: An Introduction*. Polity.

Coicaud, J. (2002). *Legitimacy and Politics: A Contribution to the Study of Political Right and Political Responsibility* (D. Curtis, Ed.). Cambridge: Cambridge University Press. doi:10.1017/CBO9780511490200

Cole, M. (2022, August 22). *What's Wrong with Technocracy? - Boston Review*. Boston Review. Retrieved July 3, 2023, from https://www.bostonreview.net/articles/whats-wrong-with-technocracy/

Cole, M., Cant, C., Ustek-Spilda, F., & Graham, M. (2022). Politics by Automatic Means? A Critique of Artificial Intelligence Ethics at Work. *Frontiers in Artificial Intelligence, 5*.

Collier, P. (2019, March 15). *France's Great Debate – how it worked*. openDemocracy. Retrieved June 28, 2023, from

https://www.opendemocracy.net/en/can-europe-make-it/frances-great-debate-how-it-worked/

Cotter, K., & Reisdorf, B.C. (2020). Algorithmic Knowledge Gaps: A New Dimension of (Digital) Inequality.

Couldry, N., & Mejias, U. A. (2019). *The costs of connection : how data is colonizing human life and appropriating it for capitalism* (Ser. Culture and economic life). Stanford University Press.

Crawford, K., & Paglen, T. (2021). Excavating AI: the politics of images in machine learning training sets. *AI & SOCIETY, 36*, pp. 1105 - 1116.

Cristianini, N., & Scantamburlo, T. (2020). On social machines for algorithmic regulation. *AI & SOCIETY, 35*, pp. 645 - 662.

Crowell, R. (2023). Why AI's diversity crisis matters, and how to tackle it. *Nature*. https://doi.org/10.1038/d41586-023-01689-4

Danks, D., & London, A.J. (2017). Algorithmic Bias in Autonomous Systems. *International Joint Conference on Artificial Intelligence.*

Davis, K. (1991). Kantian "Publicity" and Political Justice. *History of Philosophy Quarterly*, 8(4). https://philpapers.org/rec/DAVKPA

Desrosières, A. (1998). *The Politics of Large Numbers: A History of Statistical Reasoning.* Harvard University Press.

de Vries, M.J. (2018). Philosophy of Technology: Themes and Topics. In: de Vries, M. (eds) Handbook of Technology Education. Springer International Handbooks of Education. Springer, Cham. https://doi-org.proxy.library.uu.nl/10.1007/978-3-319-44687-5_1

Dewey, J. (1981). *The Later Works of John Dewey, 1925-1953* (J. A. Boydston, Ed.). SIU Press.

Dillard, K.N. (2013). Envisioning the Role of Facilitation in Public Deliberation. *Journal of Applied Communication Research, 41*, 217 - 235.

Dobbe, R., Krendl Gilbert, T., & Mintz, Y. (2021). Hard choices in artificial intelligence. Artificial Intelligence, 300, 103555. https://doi.org/10.1016/j.artint.2021.103555

Domingos, P. (2015). *The Master Algorithm: How the Quest for the Ultimate Learning Machine Will Remake Our World*. Penguin UK.

Douglas, H. E. (2009). *Science, Policy, and the Value-Free Ideal*. University of Pittsburgh Press. https://doi.org/10.2307/j.ctt6wrc78

Dryzek, J. S. & Niemeyer, S. (2011). Mini-Publics and Their Macro Consequences. In J.S. Dryzek (Ed.), *Foundations and Frontiers of Deliberative Governance* (pp. 155-176). Oxford Academic. https://doi-org.proxy.library.uu.nl/10.1093/acprof:oso/9780199562947.003.0008

Dworkin, G. (2020). Paternalism. In E. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Fall 2020 ed.). Stanford University. https://plato.stanford.edu/archives/fall2020/entries/paternalism

Elish, M.C. (2019). Moral Crumple Zones: Cautionary Tales in Human-Robot Interaction. *Engaging Science, Technology, and Society*.

Ellul J. (1964). *The technological society*. Knopf.

Eschenbach, W.J. (2021). Transparency and the Black Box Problem: Why We Do Not Trust AI. *Philosophy & Technology*.

Estlund, D. (2003). Why Not Epistocracy? In N. Reshotko (Ed.), *Desire, Identity, and Existence: Essays in Honour of T.M. Penner*. (pp. 53–70). Academic Printing and Publishing. https://doi.org/10.2307/j.ctv10kmfns.8

Eubanks, V. (2018). *Automating inequality : how high-tech tools profile, police, and punish the poor* (First). St. Martin's Press.

Fadia, A., Nayfeh, M., & Noble, J. (2020, September 16). *Follow the leaders: How governments can combat intensifying cybersecurity risks*. McKinsey & Company. Retrieved June 21, 2023, from https://www.mckinsey.com/industries/public-sector/our-insights/follow-the-leaders-how-governments-can-combat-intensifying-cybersecurity-risks

Fazelpour, S., & Danks, D. (2021). Algorithmic bias: Senses, sources, solutions. *Philosophy Compass*.

Fishkin, J. S. (2018). Democracy When the People Are Thinking. In *Oxford University Press eBooks*. https://doi.org/10.1093/oso/9780198820291.001.0001

Foucault, M. (1995). *Discipline and punish : the birth of the prison* (Second Vintage books). Vintage Books.

Fox, J. (2015). Social Accountability: What Does the Evidence Really Say? *World Development, 72*, pp. 346-361.

Fussell, S. (2020, June 24). An Algorithm That "Predicts" Criminality Based on a Face Sparks a Furor. *WIRED.* https://www.wired.com/story/algorithm-predicts-criminality-based-face-sparks-furor/

Gabriel, I. (2020). Artificial Intelligence, Values, and Alignment. *Minds and Machines, 30*, pp. 411 - 437.

Gastil, J. (2017). Designing Public Deliberation at the Intersection of Science and Public Policy. In K. H., Jamieson, D. M. Kahan, & D. A. Scheufele (Eds.), *The Oxford Handbook of the Science of Science Communication*, Oxford Library of Psychology. https://doi-org.proxy.library.uu.nl/10.1093/oxfordhb/9780190497620.013.26

Gigerenzer, G. (1991). From Tools to Theories : A Heuristic of Discovery in Cognitive Psychology. *Psychological Review*, 98, 254-267.

Gilley, B. (2017). Technocracy and democracy as spheres of justice in public policy. *Policy Sciences*, 50, pp. 9-22.

Global Trends in Government Innovation 2023. (2023). OECD. Retrieved June 3, 2023, from https://doi.org/10.1787/0655b570-en.

Gohel, P., Singh, P., & Mohanty, M. (2021). Explainable AI: current status and future directions. *ArXiv, abs/2107.07045*.

Goodin, R. E. (1985). *Protecting the Vulnerable: A Reanalysis of Our Social Responsibilities.* The University of Chicago Press. http://ci.nii.ac.jp/ncid/BA00155521

Gosepath, S. (2021). Equality. In E. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Summer 2021 ed.). Stanford University. https://plato-stanford-edu.proxy.library.uu.nl/archives/sum2021/entries/equality

Grant, J.M., Motter, L.A., & Tanis, J. (2011). Injustice at Every Turn: A Report of the National Transgender Discrimination Survey.

Green, B., & Viljöen, S. (2020). Algorithmic realism: expanding the boundaries of algorithmic thought. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*.

Green, B. (2021). Escaping the 'Impossibility of Fairness': From Formal to Substantive Algorithmic Fairness. InfoSciRN: Machine Learning (Sub-Topic).

Grim, P., Singer, D.J., Bramson, A.L., Holman, B., Mcgeehan, S., & Berger, W.J. (2019). Diversity, Ability, and Expertise in Epistemic Communities. *Philosophy of Science, 86*, pp. 98–123.

Gutmann, A. & Thompson, F. D. (2018). Reflections on Deliberative Democracy: When Theory Meets Practice. In A. Bächtiger et al. (Eds.), *The Oxford Handbook of Deliberative Democracy*, Oxford Handbooks. https://doi-org.proxy.library.uu.nl/10.1093/oxfordhb/9780198747369.013.44

Habermas, J. (1996). Between Facts and Norms. In *The MIT Press eBooks*. https://doi.org/10.7551/mitpress/1564.001.0001

Harju, B. (2019). Too Much Information: Self-Monitoring and Confessional Culture. In Zappe, F., & Gross, A. (2019). Introduction. In: Zappe, F., & Gross, A. (Eds.), *Surveillance - society - culture* (Ser. Contributions to english and american literary studies (ceals), vol. 3). Peter Lang.

Hartz-Karp, J., Carson, L., & Briand, M. (2018). Deliberative Democracy as a Reform Movement. In A. Bächtiger et al. (Eds.), *The Oxford Handbook of Deliberative Democracy*, Oxford Handbooks. https://doi-org.proxy.library.uu.nl/10.1093/oxfordhb/9780198747369.013.41

Hayward, C.R. (2021). Why does publicity matter? Power, not deliberation. *Journal of Political Power, 14*, pp. 176 - 195.

Hedden, B. (2021). On statistical criteria of algorithmic fairness. *Philosophy & Public Affairs, 49*, pp. 209-231.

Henderson, L. (2022). The Problem of Induction. In E. Zalta & U. Nodelman (Eds.), *The Stanford Encyclopedia of Philosophy* (Winter 2022 ed.). Stanford University. https://plato.stanford.edu/archives/win2022/entries/induction-problem

Heikkilä, M. (2022, April 13). Dutch scandal serves as a warning for Europe over risks of using algorithms. *POLITICO*. https://www.politico.eu/article/dutch-scandal-serves-as-a-warning-for-europe-over-risks-of-using-algorithms/

Henley, J. (2021, January 15). Dutch government faces collapse over child benefits scandal. *The Guardian*. https://www.theguardian.com/world/2021/jan/14/dutch-government-faces-collapse-over-child-benefits-scandal

Hidalgo, C. (2018). *A bold idea to replace politicians* [Video]. TED Talks. https://www.ted.com/talks/cesar_hidalgo_a_bold_idea_to_replace_politicians?language=en

Hildebrandt, M. (2008). Defining Profiling: A New Type of Knowledge? In M. Hildebrandt & S. Gutwirth (Eds.), *Profiling the European Citizen*. Springer. https://link-springer-com.proxy.library.uu.nl/book/10.1007/978-1-4020-6914-7

Hong, L., & Page, S.E. (2004). Groups of diverse problem solvers can outperform groups of high-ability problem solvers. *Proceedings of the National Academy of Sciences of the United States of America, 101 46*, 16385-9 .

Honneth, A. (2001). Recognition or redistribution? *Theory, Culture & Society*, *18*(2-3), pp. 43–55. https://doi.org/10.1177/02632760122051779

Huang, B., Thomas, T., Groenewolt, B., Claasen, Y., & Berkum, E.C. (2021). Effectiveness of incentives offered by mobile phone app to encourage cycling: A long-term study. *IET Intelligent Transport Systems*.

Human rights by design: future-proofing human rights protection in the era of AI. (2023). Council of Europe Commissioner for Human Rights. Retrieved June 2, 2023, from https://www.coe.int/en/web/commissioner/thematic-work/artificial-intelligence

Jacquet, V. (2018). The Role and the Future of Deliberative Mini-publics: A Citizen Perspective. *Political Studies, 67*, pp. 639–657.

Jankin, S., Esteve, M., & 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*.

Kahneman, D. (2011). *Thinking, Fast and Slow*. Penguin UK.

Katz, Y. (2020). *Artificial whiteness : politics and ideology in artificial intelligence*. Columbia University Press.

Kearns, M., & Roth, A. (2019). The Ethical Algorithm: The Science of Socially Aware Algorithm Design. Oxford University Press.

Keyes, O. (2018). The Misgendering Machines. *Proceedings of the ACM on Human-computer Interaction*, 2(CSCW), pp. 1–22. https://doi.org/10.1145/3274357

Kirichenko, L., Radivilova, T., & Carlsson, A. (2018). Detecting cyber threats through social network analysis: short survey. *ArXiv, abs/1805.06680*.

Kleinberg, J.M., Mullainathan, S., & Raghavan, M. (2016). Inherent Trade-Offs in the Fair Determination of Risk Scores. *ArXiv, abs/1609.05807*.

Kleinberg, J., Ludwig, J., Mullainathan, S., & Rambachan, A. (2018). Algorithmic Fairness. *AEA Papers and Proceedings*, 108, pp. 22–27. https://doi.org/10.1257/pandp.20181018

Kuppler, M., Kern, C., Bach, R.L., & Kreuter, F. (2021). Distributive Justice and Fairness Metrics in Automated Decision-making: How Much Overlap Is There? *ArXiv, abs/2105.01441*.

Kurzweil, R. (2005). *The Singularity Is Near: When Humans Transcend Biology*. Penguin Books.

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

Lafont, C. (2015). Deliberation, Participation, and Democratic Legitimacy: Should Deliberative Mini-Publics Shape Public Policy? *Journal of Political Philosophy, 23*, pp. 40-63.

Lam, T. (2021). The People's Algorithms: Social Credits and the Rise of China's Big (Br)other. In Mennicken, A., & Salais, R. (Eds.), *The New Politics of Numbers: Utopia, Evidence and Democracy.* Palgrave Macmillan.

Lambrecht, A., & Tucker, C. (2019). Algorithmic Bias? An Empirical Study of Apparent Gender-Based Discrimination in the Display of STEM Career Ads. *Manag. Sci., 65*, 2966-2981.

Landemore, H. (2013). Democratic reason: politics, collective intelligence, and the rule of the many. *Choice Reviews Online*, *51*(01), 51–0525. https://doi.org/10.5860/choice.51-0525

Landemore, H. (2020). Open Democracy. In *Princeton University Press eBooks*. https://doi.org/10.23943/princeton/9780691181998.001.0001

Landemore, H. (2022). Can AI bring deliberative democracy to the masses? In *NYU Law* [Presentation]. Colloquium in Legal, Political, and Social Philosophy, United States of America. https://www.law.nyu.edu/centers/lawphilosophy/colloquium

Lau, T. (2020, April 1). *Predictive policing explained*. Brennan Center for Justice. https://www.brennancenter.org/our-work/research-reports/predictive-policing-explained

Lee, M.S., Floridi, L., & Singh, J. (2020). Formalising trade-offs beyond algorithmic fairness: lessons from ethical philosophy and welfare economics. *AI and Ethics, 1*, pp. 529 - 544.

Leino, M., Kulha, K., Setälä, M., & Ylisalo, J. (2022). Expert hearings in mini-publics: How does the field of expertise influence deliberation and its outcomes? *Policy Sciences, 55*, pp. 429 - 450.

Leins, K., & Kaspersen, A. (2021, November 10). Seven Myths of Using the Term "Human on the Loop": "Just What Do You Think You Are Doing, Dave?" *Carnegie Council for Ethics in*

International Affairs.
https://www.carnegiecouncil.org/media/article/7-myths-of-using-the-term-human-on-the-loop

Leufer, D. (2023). Computers are binary, people are not: how AI systems undermine LGBTQ identity. *Access Now*.
https://www.accessnow.org/how-ai-systems-undermine-lgbtq-identity/

Loi, M., & Heitz, C. (2022). Is calibration a fairness requirement?: An argument from the point of view of moral philosophy and decision theory. *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*

Lyon, D. (2001). *Surveillance Society: Monitoring Everyday Life*. McGraw-Hill Education (UK).

Madan, R., & Ashok, M. (2022). AI adoption and diffusion in public administration: A systematic literature review and future research agenda. *Gov. Inf. Q., 40*, 101774.

Markie, P., & Folescu, M. (2023). Rationalism vs. Empiricism. In E. Zalta & U. Nodelman (Eds.), *The Stanford Encyclopedia of Philosophy* (Spring 2023 ed.). Stanford University.
https://plato.stanford.edu/archives/spr2023/entries/rationalism-empiricism

McQuillan, D. (2020). "The Political Affinities of AI". In: Sudmann, A. (Ed.). (2020). *The democratization of artificial intelligence : net politics in the era of learning algorithms* (Ser. Ki-kritik, 1). transcript-Verlag. https://doi.org/10.14361/9783839447192

McQuillan, D. (2022). *Resisting AI: An Anti-fascist Approach to Artificial Intelligence*. Policy Press.

Medaglia, R., Gil-Garcia, J.R., & Pardo, T.A. (2021). Artificial Intelligence in Government: Taking Stock and Moving Forward. *Social Science Computer Review, 41*, pp. 123 - 140.

Medaglia, R., & Tangi, L. (2022). The adoption of Artificial Intelligence in the public sector in Europe: drivers, features, and impacts. *Proceedings of the 15th International Conference on Theory and Practice of Electronic Governance*.

Merriam-Webster. (n.d.). Algorithm. In Merriam-Webster.com dictionary. Retrieved July 9, 2023, from https://www.merriam-webster.com/dictionary/algorithm

Merriam-Webster. (n.d.). Subjectivism. In Merriam-Webster.com dictionary. Retrieved July 8, 2023, from https://www.merriam-webster.com/dictionary/subjectivism

Mheslinga. (2022, June 1). Behavioral economics, explained. *University of Chicago News*.
https://news.uchicago.edu/explainer/what-is-behavioral-economics

Mikhailenko, O.V., & Dorrer, G.A. (2020). Methodological problems of big data and artificial intelligence in the medical specialists training. *Journal of Physics: Conference Series, 1691*.

Miles-Johnson, T. (2015). "They Don't Identify With Us": Perceptions of Police by Australian Transgender People. *International Journal of Transgenderism, 16*, pp. 169 - 189.

Miller, B. (2021). Government Chatbots Now a Necessity for States, Cities, Counties. *GovTech.* https://www.govtech.com/products/government-chatbots-now-a-necessity-for-states-cities-counties.html

Miller, D. (2021). Justice. In E. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Fall 2021 ed.). Stanford University. https://plato.stanford.edu/archives/fall2021/entries/justice

Mission. (n.d.). Humanity+. https://www.humanityplus.org/about

Mitchell, S., Potash, E., Barocas, S., D'Amour, A., & Lum, K. (2021). Algorithmic Fairness: Choices, Assumptions, and Definitions. *Annual Review of Statistics and Its Application*.

Moser, C., Den Hond, F., & Lindebaum, D. (2022). What Humans Lose When We Let AI Decide. *MIT Sloan Management Review*, 63(3), pp. 12-14. https://sloanreview.mit.edu/article/what-humans-lose-when-we-let-ai-decide/

Neblo, M. A. (2015). *Deliberative democracy between theory and practice*. Cambridge University Press.

Neslen, A. (2021, October 20). FEATURE-Pushback against AI policing in Europe heats up over racism fears. *U.S*. https://www.reuters.com/article/europe-tech-police-idINL8N2R92HQ

New Poll: Public fears over government use of Artificial Intelligence. (n.d.). ECNL. https://ecnl.org/news/new-poll-public-fears-over-government-use-artificial-intelligence

Nguyen, V.B., Schlötterer, J., & Seifert, C. (2022). Explaining Machine Learning Models in Natural Conversations: Towards a Conversational XAI Agent. *ArXiv, abs/2209.02552*.

Obermeyer, Z., Powers, B.W., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science, 366*, pp. 447 - 453.

OECD's live repository of AI strategies & policies - OECD.AI. (n.d.). Retrieved June 2, 2023, from https://oecd.ai/en/dashboards/overview

O'Reilly, T. (2013). Open data and algorithmic regulation. In Goldstein, B., & Dyson, L. (Eds.), *Beyond transparency: open data and the future of civic innovation.* Code for America Press, San Francisco, pp. 289–300

O'Riordan, T., & Cameron, J. D. (2013). Interpreting the Precautionary Principle. In *Routledge eBooks*. https://doi.org/10.4324/9781315070490

Ossewaarde, M., & Gulenc, E. (2020). National Varieties of Artificial Intelligence Discourses: Myth, Utopianism, and Solutionism in West European Policy Expectations. *Computer, 53*, pp. 53-61.

Pannett, R. (2022, August 30). France uses AI to spot (and tax) undeclared swimming pools. *Washington Post.* https://www.washingtonpost.com/world/2022/08/30/france-undeclared-swimming-pools-artificial-intelligence/

Pareek, V. (2019). Non-binary Gender and Data. In *Handbook of Gender and Open Data*. https://cis.pubpub.org/pub/non-binary-gender-data

Pearl, J., Glymour, M. R. K., & Jewell, N. P. (2016). Causal inference in statistics : a primer. In *John Wiley & Sons, Inc. eBooks.* http://perpus.univpancasila.ac.id/index.php?p=show_detail&id=124976

Peter, F. (2007). Democratic legitimacy and proceduralist social epistemology. *Politics, Philosophy & Economics, 6*, pp. 329 - 353.

Peter, F. (2009). *Democratic Legitimacy*. Routledge.

Peter, F. (2017). Political Legitimacy. In E. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Summer 2017 ed.). Stanford University. https://plato.stanford.edu/archives/sum2017/entries/legitimacy

Piedmont, R.L. (2014). Bias, Statistical. In A.C. Michalos (Ed.), *Encyclopedia of Quality of Life and Well-Being Research*. Springer, Dordrecht. https://doi-org.proxy.library.uu.nl/10.1007/978-94-007-0753-5_2865

Piloto, C. (2023, March 13). The Gender Gap in STEM: Still Gaping in 2023. *MIT Professional Education.* Retrieved June 6, 2023, from https://professionalprograms.mit.edu/blog/leadership/the-gender-gap-in-stem/

Pitt, J. C. (2000). *Thinking about Technology: Foundations of the Philosophy of Technology.* Seven Bridges Press.

Pitt, J. C. (2014). "Guns Don't Kill, People Kill"; Values in and/or Around Technologies. In: Kroes, P., Verbeek, PP. (eds), *The Moral Status of Technical Artefacts. Philosophy of Engineering and Technology*, vol 17. Springer, Dordrecht. https://doi-org.proxy.library.uu.nl/10.1007/978-94-007-7914-3_6

Poel, I.V. (2020). Three philosophical perspectives on the relation between technology and society, and how they affect the current debate about artificial intelligence. *Human Affairs, 30*, pp. 499 - 511.

Pozen, D.E. (2020). Seeing Transparency More Clearly. *Public Administration Review, 80*, 326-331.

Pumperla, M., & Ferguson, K. (2019). *Deep Learning and the Game of Go* (1st ed.). Manning Publications.

Rambachan, A., & Roth, J. (2019). Bias In, Bias Out? Evaluating the Folk Wisdom. *ArXiv, abs/1909.08518*.

Rawls, J. (2009). *A Theory of Justice*. Harvard University Press.

Romeijn, J. (2011). Statistics as Inductive Inference. In *Elsevier eBooks* (pp. 751–774). https://doi.org/10.1016/b978-0-444-51862-0.50024-1

Rosenberg, M., Confessore, N., & Cadwalladr, C. (2019, March 19). How Trump Consultants Exploited the Facebook Data of Millions. *The New York Times*. https://www.nytimes.com/2018/03/17/us/politics/cambridge-analytica-trump-campaign.html

Rostboll, C.F. (2008). Deliberative Freedom : Deliberative Democracy As Critical Theory. SUNY Press.

Russell, S., & Norvig, P. (2021). *Artificial Intelligence: A Modern Approach* (4th ed.). Pearson.

Sajid, H. (2023, April 20). 7 Practical Applications of AI in Government. *V7.* https://www.v7labs.com/blog/ai-in-government

Santana, M.C., De Marsico, M., Nappi, M., & Riccio, D. (2017). MEG: Texture operators for multi-expert gender classification. *Comput. Vis. Image Underst., 156*, 4-18.

Scheuerman, M.K., Paul, J.M., & Brubaker, J.R. (2019). How Computers See Gender: An Evaluation of Gender Classification in Commercial Facial Analysis and Image Labeling Services. *Proceedings of the ACM on Human-computer Interaction,* 3(CSCW), pp. 1-33. https://doi.org/10.1145/3359246

Schwartz, O. (2018, October 31). "The discourse is unhinged": how the media gets AI alarmingly wrong. *The Guardian.* https://www.theguardian.com/technology/2018/jul/25/ai-artificial-intelligence-social-media-bots-wrong

Sherover, C. M. (1992). The Conditions of Freedom: A New World Order. *Public Affairs Quarterly*, *6*(4), pp. 415–433. http://www.jstor.org/stable/40435824

Shneiderman, B. (2021). Human- Centered AI: Computer scientists should build devices to enhance and empower--not replace--humans. *Issues in Science and Technology*, *37*(2), 56+.

Shrader-Frechette, K.S. (1994). Reductionist Philosophy of Technology: Stones Thrown from Inside a Glass House. *Techné: Research in Philosophy and Technology, 5*, pp. 21-28.

Sien, J.P., Lim, K., & Au, P. (2019). Deep Learning in Gait Recognition for Drone Surveillance System. *IOP Conference Series: Materials Science and Engineering, 495*.

Simons, J., & Frankel, E. (2023, February 21). *Why democracy belongs in artificial intelligence*. Princeton University Press. Retrieved June 8, 2023, from https://press.princeton.edu/ideas/why-democracy-belongs-in-artificial-intelligence

Siu, A. (2022). Using an AI-assisted deliberation platform to achieve deliberative democracy. In *ACM Collective Intelligence Conference* [Video]. ACM Collective Intelligence Conference 2022 (CI 2022). https://www.youtube.com/watch?v=CEKKHi-feC0

Sorial, S. (2022). Deliberation and the Problems of Exclusion and Uptake: The Virtues of Actively Facilitating Equitable Deliberation and Testimonial Sensibility. *Ethical Theory and Moral Practice, 25*, pp. 215 - 231.

Sousa, W.G., Melo, E.R., Bermejo, P.H., Farias, R.A., & Gomes, A.D. (2019). How and where is artificial intelligence in the public sector going? A literature review and research agenda. *Gov. Inf. Q., 36*.

Stelmaszak, M. (2021). How To Train Your Algo: Investigating the Enablers of Bias in Algorithmic Development. *International Conference on Interaction Sciences*.

Suresh, H., & Guttag, J.V. (2019). A Framework for Understanding Unintended Consequences of Machine Learning. *ArXiv, abs/1901.10002.*

Swer, G. M. (2014). Determining technology: myopia and dystopia. *South African Journal of Philosophy, 33(2),* 201–210. https://doi.org/10.1080/02580136.2014.923696

Swer, G.M., & du Toit, J. (2020). A Manifesto for Messy Philosophy of Technology: The History and Future of an Academic Field. *Teorie vědy / Theory of Science*.

Sætra, H.S. (2020). A shallow defence of a technocracy of artificial intelligence: Examining the political harms of algorithmic governance in the domain of government. *Technology in Society, 62*, 101283 - 101283.

Taylor, C. (1979). What's Wrong with Negative Liberty. In A. Ryan (Ed.), *The Idea of Freedom*. Oxford: Oxford University Pres. pp. 175–93.

Thaler, R. H., & Sunstein, C. R. (2008). Nudge: Improving Decisions about Health, Wealth, and Happiness. Yale University Press.

Tversky, A. , & Kahneman, D. (1974). Judgment under Uncertainty : Heuristics and Biases. *Science*, 185, pp. 1124-1131.

UN (United Nations). (1948). *Universal Declaration of Human Rights.* https://www.un.org/en/about-us/universal-declaration-of-human-rights

Vaidhyanathan, S. (2012). *The Googlization of Everything: (And Why We Should Worry).* Univ of California Press.

Valle-Cruz, D., Gómez, E.A., Sandoval-Almazán, R., & Criado, J.I. (2019). A Review of Artificial Intelligence in Government and its Potential from a Public Policy Perspective. *Proceedings of the 20th Annual International Conference on Digital Government Research*.

van Dijck, G. (2022). Predicting Recidivism Risk Meets AI Act. *European Journal on Criminal Policy and Research, 28*, pp. 407 - 423.

van Giffen, B., Herhausen, D., & Fahse, T.B. (2022). Overcoming the pitfalls and perils of algorithms: A classification of machine learning biases and mitigation methods. *Journal of Business Research*.

Verbeek, PP. (2020). Politicizing Postphenomenology. In Miller, G., & Shew, A. (Eds.), *Reimagining Philosophy and Technology, Reinventing Ihde*. Springer.

Verbeek, PP. (2022). The Empirical Turn. In S. Vallor (Ed.), *The Oxford Handbook of Philosophy of Technology* (pp. 35–54). essay, Oxford University Press.

Viola, L. A., & Laidler, P. (Eds.). (2022). *Trust and transparency in an age of surveillance* (Ser. Routledge studies in surveillance). Routledge.

Volpicelli, G. (2023, March 6). ChatGPT broke the EU plan to regulate AI. *POLITICO*. https://www.politico.eu/article/eu-plan-regulate-chatgpt-openai-artificial-intelligence-act/

Wang, Y., Sun, T., Li, S., Yuan, X., Ni, W., Hossain, E., & Poor, H.V. (2023). Adversarial Attacks and Defenses in Machine Learning-Powered Networks: A Contemporary Survey. *ArXiv, abs/2303.06302*.

Werkheiser, I. (2016). Community Epistemic Capacity. *Social Epistemology, 30*, pp. 25 - 44.

Westen, D. (2008). *The Political Brain: The Role of Emotion in Deciding the Fate of the Nation.* Public Affairs.

Winner, L. (1977). *Autonomous Technology: Technics-out-of-Control as a Theme in Political Thought*. MIT Press.

Wirtz, B.W., & Müller, W.M. (2018). An integrated artificial intelligence framework for public management. *Public Management Review, 21*, pp. 1076 - 1100.

Wittkower, D.E. (2018). "Discrimination". In: Pitt, J. C., & Shew, A. (Eds.). (2018). *Spaces for the future : a companion to philosophy of technology.* Routledge.

Yan, W. Q. (2016). *Introduction to Intelligent Surveillance*. Springer.

Yeung, K. (2022). The New Public Analytics as an Emerging Paradigm in Public Sector Administration. *Tilburg Law Review.*

Young, I. (1996). Six Communication and the Other: Beyond Deliberative Democracy. In S. Benhabib (Ed.), *Democracy and Difference: Contesting the Boundaries of the Political* (pp. 120-136). Princeton: Princeton University Press. https://doi-org.proxy.library.uu.nl/10.1515/9780691234168-007

Zappe, F. (2019). Gazing Back at the Monster. In: Zappe, F., & Gross, A. (Eds.), *Surveillance - society - culture* (Ser. Contributions to english and american literary studies (ceals), vol. 3). Peter Lang.

Zappe, F., & Gross, A. (2019). Introduction. In: Zappe, F., & Gross, A. (Eds.), *Surveillance - society - culture* (Ser. Contributions to english and american literary studies (ceals), vol. 3). Peter Lang.

Zheng, S., Trott, A.R., Srinivasa, S., Naik, N.V., Gruesbeck, M., Parkes, D.C., & Socher, R. (2020). The AI Economist: Improving Equality and Productivity with AI-Driven Tax Policies. *ArXiv, abs/2004.13332.*