# Why Algorithmic Decision Making Is Not Value-Neutral

## Proposing a taxonomy of values in algorithmic decision making

Master's thesis to obtain a MSc degree in Artificial Intelligence

by

Marijn Biekart
Student number: 9885234

June 21, 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. Michael De

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

## Abstract

Algorithmic decision making (ADM) is used to assist decisions that have far-reaching consequences for individuals and society as a whole, for example in hiring and criminal law. As such, it is important that ADM is fair. It is commonly believed that ADM is objective and neutral, which supports its fairness. The first aim of this thesis is to show that ADM is not, and cannot be, objective or neutral. Instead, ADM necessarily contains value judgments. In order to prove the existence of values in ADM, this thesis uses arguments from the philosophy of science, that show that *science* is not value-free. The parallels and differences between science and ADM indicate that values play an even bigger role ADM, compared to science. The second aim of this thesis is to propose a taxonomy of values in ADM, which indicates *where* values play a role, *what* values play a role, and *how* they play a role. There are two main purposes of the taxonomy: (1) it can be used by developers and regulators to recognize the values that play a role in ADM systems, ideally resulting in less unintentional outcomes; (2) it can be used to regulate ADM by informing public sector policies and laws. The practical use of the taxonomy is demonstrated by a case study, focusing on Rotterdam's welfare fraud detection system, which uses risk profiles to indicate which recipients have a higher risk of committing fraud. This thesis provides a deeper understanding of the relation between values, bias, and unfairness in ADM. By acknowledging that ADM cannot be value-neutral, this thesis shifts the focus from omitting bias to managing bias, in an effort to make ADM fairer for everyone.

**Keywords:** algorithmic decision making, value judgments, fair AI, philosophy of AI, ethics, risk profiling.

# Contents

# 1 Introduction

The research goal of artificial intelligence (AI) can be described in many different ways. A widely accepted definition can be found in Russell & Norvig (2010), who recognize four main goals of AI. With a *human-based approach*, the aim of AI is either to build systems that think like humans (1), or systems that act like humans (2). However, if one has an *ideal rationality* approach, the aim of AI is either to build systems that think rationally (3) or act rationally (4), like an 'ideal' human would (Russell & Norvig, 2010). Different AI applications have different aims. For example, a robot that is designed to take care of elderly people should arguably act like a human would (human-based approach). Conversely, a system that helps decide on a convict's sentence length should act rationally (ideal rationality approach).

This last case is an example of *algorithmic decision making* (ADM). ADM uses outputs produced by algorithms to make or assist decisions. Another example of ADM systems can be found in hiring procedures: ADM systems can be used to predict the chances that an applicant will succeed in a particular job. This prediction can be used to make a decision about hiring the applicant. ADM systems can either make a decision without a human in the loop, or they can be used to *assist* humans by advising them what to decide. This thesis will focus on ADM systems in the latter sense, where they are used as an advisory tool to assist human decision makers.

Together with today's vast availability of data (Big Data), ADM can be used to tackle complex problems for researchers, companies, governments, and other actors in the public sector (Lepri et al., 2018). As such, ADM is used to assist decisions that have far-reaching individual or societal implications, such as hiring, criminal sentencing, and fraud detection. Traditionally, decisions in these domains have been made by humans, often experts. However, human decision making has often shown to have serious limitations and bias, resulting in inefficient and/or unjust processes and outcomes (Fiske, 1998). Lepri et al. (2018) describe the turn towards ADM with Big Data as a paradigm shift, reflecting "the demand for greater objectivity, evidence-based decision making, and a better understanding of our individual and collective behaviors and needs" (p. 612).

Because ADM systems are used in settings that affect real people, it is important that they behave fairly. There is a common belief that objectivity and evidence-based approaches are important factors for fair predictions. Although rarely explicitly articulated, the argument for this belief seems to be as follows: objective approaches are not influenced by values or value judgments and therefore they cannot be fair or unfair. A human decision maker has their own prejudice, their own values that they find important. This may compromise their ability to make decisions that are fair, although not every biased decision is driven by malevolence. For example, it has been shown that in-group bias (i.e., favoring someone from one's own group over someone from outside one's own group) is actually pre-coded in our brain (Molenberghs, 2013). This finding suggests that, even if one were aware of their biases, it would be difficult (if not impossible) to avoid being influenced by them.

Contrary to humans, ADM systems are thought to have no evolutionary neural biases or value judgments of their own, so the common belief is that this should make them more capable of making fair decisions. Furthermore, the turn towards Big Data reflects a common belief that objective, evidence-based data provide a good starting point for fair decision-making. Large data sets are seen as objective bodies of knowledge, and we use them to analyze human behavior and support our decision-making (e.g., to help predict if someone has an increased risk of heart failure). Data merely are seen as 'representations of the world as it is', again reflecting no values or personal and subjective judgments.

Despite the promising objective and evidence-based approach of ADM, recent research has shown that systems often do not make fair predictions (Crawford & Paglen, 2021; Allhutter et al., 2020; Angwin et al., 2016)). For example, Angwin et al. (2016) find that the COMPAS system, which predicts the recidivism risks of convicts, is biased against people of color. Findings like these gave rise to the field of *fair AI*, which draws attention to the ways in which algorithms can be unfair and how their unfairness can be improved (e.g., Lepri et al. (2018); Binns (2018); Langenkamp et al. (2020)).

Fair AI often focuses on *bias* in data sets, stating that data sets are not fair to begin with. This is a logical approach, because 'the world as it is' is not fair: there are unfair inequalities between different genders, races, sexual orientations and socioeconomic classes, to name a few. Inevitably, these inequalities are reflected in existing data and fed into ADM systems. Because ADM systems are 'objective' and cannot reason about what is fair and unfair, the predictions they make are only reinstating the unfairness in the data. For example, it has historically been more difficult for women to get into leadership positions than for men (Perez, 2019). This inequality is reflected in data about the distribution of male versus female leaders. From this data, an algorithm that aids hiring decisions might learn that men are better leaders, such that it only selects male applicants. As a result of the biased algorithm, more men get into leadership positions, making the distribution between men and women in leadership positions even more unequal.

There is an intuitive feeling that bias in ADM is directly related to the value-ladenness of systems and fairness. Yet, this connection is often not made explicit, and there is a slight difference between these concepts. My aim in this thesis is to add to the research field of fair AI, by explaining the connection between objectivity, values, bias and fairness. As we will see, objectivity is often described as 'the absence of value judgments', 'value-freedom' or 'value-neutrality'. This definition reveals the direct (negative) connection between objectivity and values. In turn, values are often associated with bias or prejudice, because they represent personal and subjective beliefs that someone may have. For example, if someone highly values self-determination, they might be biased to decide against mandatory vaccination programs. If someone values the lives of certain (groups of) people more than others, because they were raised in societies were this has been the case throughout history, they become biased against marginalized people. Thus, values are an important motivator behind bias, which often results in unfair outcomes for marginalized groups. Following this line of reasoning, it seems logical to conclude that objectivity leads to fairness: values and biases are left behind in an objective approach.

With the current rise of fair AI, the conviction that ADM is automatically fair because there are no values involved, is being questioned. In order to say more about the status of values in ADM, it is useful to return to a field in which objectivity is particularly important: science. Just like ADM, the natural and social sciences (science hereafter) are often claimed to be objective and value-free. There is an important sub-field of the philosophy of science that is dedicated to objectivity in science (e.g., see Reiss & Sprenger (2020)). Scientific objectivity can be defined in different ways. In this thesis, objectivity will be conceptualized as *absence of normative commitments and value-freedom*, as defined by Lacey (1999). I will abbreviate this conception as *value-freedom.* A good scientific theory should not be dependent on the scientist's personal values and beliefs, but should only hold because of factual and verifiable evidence.

The ideal of value-free science is criticized by different authors (Douglas, 2009; Lacey, 1999; Anderson, 2004; Longino, 1995), either for not being attainable or for not being desirable. The main point of these authors is to show the different ways in which science is not, and possible cannot be, value-free. Furthermore, they show what values *do* play a role in science. By doing this, they create awareness of certain things or groups that have been overlooked in science. Furthermore, they show how value judgments can be *managed*, instead of omitted.

My aim in this thesis is to prove that ADM is not, and cannot be, value-neutral. I will do this by analyzing the arguments that speak against the value-neutrality of science, and investigating to what extent they can be used to prove that ADM systems are not value-neutral. Based on the debate about scientific objectivity, I will move the topic of fair ADM from a debate about *bias* to a debate about *values*. I have formulated three **research questions** to guide this thesis:

1. How does the value-neutrality of science relate to the claimed objectivity of algorithmic decision making?
2. To what extent can the philosophical arguments *against* value-neutrality of science be applied in the context of the objectivity of algorithmic decision making?
3. *Where* in the machine learning pipeline do values play a role in algorithmic decision making, *what* values may play a role, and *how* do they play a role?

To answer the last question, I will provide a taxonomy of values in algorithmic decision making. This taxonomy will specify *where* in the AI pipeline value judgments can be located, *what* value judgments play a role, and *how* these values play a role in AI and the application of AI systems. The practical use of the taxonomy will be demonstrated with a case study, focusing on Rotterdam's welfare fraud detection algorithm.

The key contribution of this thesis is to provide a deeper understanding of bias and unfairness in ADM. By acknowledging that ADM cannot be value-neutral, this thesis shifts the focus on managing bias, instead of omitting bias. Understanding the complex interactions between objectivity, neutrality, bias and fairness will help us determine in what contexts ADM should and should not be applied. Furthermore, I propose a taxonomy of values in ADM, which is currently missing in the debate around fair AI. The contribution of the taxonomy is twofold: (1) it can be used by developers and regulators to recognize the values that play a role in AI systems and it raises consciousness about the use of ADM systems, ideally resulting in less unintentional outcomes; (2) it can accordingly be used to regulate ADM by informing public sector policies and lawmakers. If certain value judgments in ADM cannot be removed, it is at least helpful to be aware of them and manage them adequately.

# 2 Theoretical Background

This section discusses previous research that is relevant for my thesis. In Section 2.1, I will explain what algorithmic decision making (ADM) is and I will demonstrate its practical use with a few recent examples. Secondly, Section 2.2 reveals how ADM systems have been shown to behave biased in a way that harmfully discriminates individuals or social groups, resulting in unfair decisions. Section 2.2 also discusses approaches that have been undertaken in an attempt to make ADM systems fairer. Finally, in Section 2.3, I will explain how the ideal of objectivity in science is criticized in the philosophy of science, often from a feminist standpoint.

## 2.1 Algorithmic Decision Making

In general, algorithmic decision making (ADM) means using outputs produced by algorithms to make or assist decisions. One of the earliest practical applications of ADM can be found in the American legal system, where algorithms are used to assist judges in determining the sentence length for a suspect. These algorithms take the details of the crime and suspect (e.g., severity of the crime, past record of the suspect) into account and perform a statistical-based mathematical operation, in order to recommend a sentence length.

ADM often makes use of supervised machine learning, which is a commonly used method in AI. Supervised machine learning algorithms use data sets that function as “ground truth”. This means that the algorithm learns with examples where the ground truth is given, allowing the algorithm to create connections between data features and their ground truths. These connections are then used to create outcomes for new instances, for which the ground truth is unknown to the algorithm.

For instance, take a supervised learning algorithm that predicts a prospective university student’s chance of success, measured in their GPA score. The predictions of this algorithm are based on data points which represent previously enrolled students and their GPA scores. Students in the data set are described along different characteristics, such as gender, age, high school grades and extracurricular activities. In ADM and machine learning, these characteristics are called *features*. Supervised machine learning algorithms learn relations between students’ features and their label (i.e., GPA score). For example, an algorithm might learn that extracurricular activities are a strong indicator of high GPA scores. To give advice about prospective students, algorithms compare the features of prospective students with the features of the students in the data set, and make an “informed” prediction about the chances of success for the prospective students.

A detailed technical explanation of ADM is beyond the scope of this thesis. However, a useful semi-technical description of the processes involved can be found in Friedler et al. (2016). They describe ADM as a mapping from a **construct space** to a **decision space**. The construct space consists of the features that the decision-maker ideally wants to take into account when making the decision. The decision space consists of the decision outcomes. For example, when assessing recidivism risk of an offender, the construct space might consist of the offender’s inclination to commit a crime (Friedler et al., 2016). The decision space consists of the decision whether to grant the offender parole.

What complicates the matter is that an individual’s features in construct space cannot be observed directly. An offender’s inclination to commit crime cannot be measured in a straightforward way. Instead, features in the construct space can only be measured indirectly, through the **observed** space. The observed space consists of proxy variables that were used in the training data

(Binns, 2018). To return to our example, the observed space might consist of facts about the offender's level of education and job occupation. Whereas inclination to commit crime is not directly observable or measurable, education and job occupation are. These features act as a proxy variable to infer the offender's inclination to commit a crime. A schematic overview of the different spaces in ADM systems can be found in Figure 1.

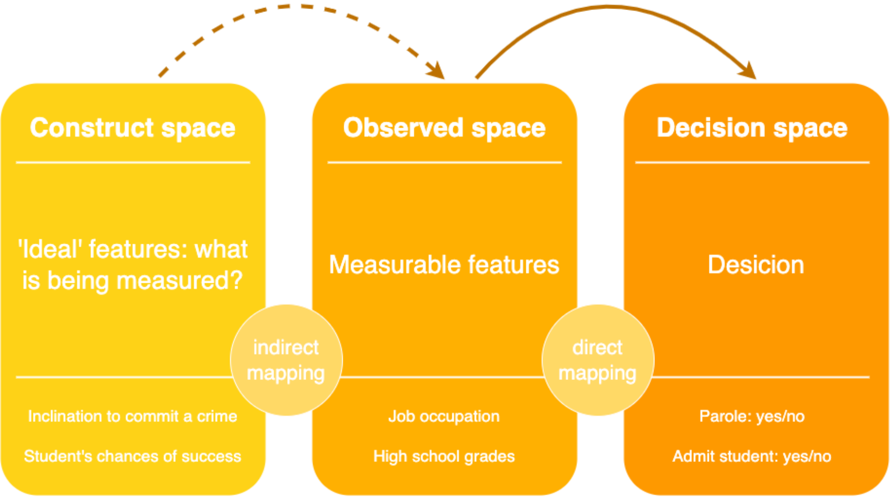

Figure 1: A schematic representation of the different spaces in ADM.

While it is theoretically possible to let algorithms make important decisions on their own, it is not often the case in practice (yet) (Danaher, 2016). ADM systems are currently generally used as algorithmic decision *support* systems (as will be demonstrated in the examples below). This means that there is a human in the loop: a professional who can assess the outcome of the system, and make an informed decision to follow or disregard its advice[1]. One reason why ADM has a more supportive role than a conclusive one, is that there are strict legal rules around the use of algorithmic decision making. It is theoretically possible to let important decisions about an individual be made by an ADM system, but it brings along quite a few legal regulations concerning individuals' rights to privacy, explanation, and objection[2]. This makes the use of algorithmic support systems more widespread than independent ADM.

For this reason, the focus in this thesis will lie on supportive ADM (simply denoted by 'ADM' hereafter). Even if legal restrictions did not prevent independent ADM systems to be applied, the focus on supportive ADM could still be justified by the fact that some of the most well-discussed ADM systems in literature play a supportive role (e.g., Angwin et al. (2016)). Furthermore, an independent ADM system deciding over individuals' lives would arguably raise more ethical questions than when ADM only plays an supportive role. Seeing supportive ADM as 'ADM light' and autonomous ADM as 'ADM plus', will strengthen my argument: if values play a role in ADM light,

[1] When talking about algorithmic decision support, the outcome of an ADM system can be seen as 'advice', which humans can choose to either ignore or follow.

[2] European legislation can be found in the GDPR (Article 4(4) and 22, Recitals (71) and (72)) and the Guidelines on Automated individual decision-making and Profiling for the purposes of Regulation 2016/679.

they will also play a role in ADM plus, and their impact will arguably be even bigger.

To end this explanation of ADM, I will provide a few examples of ADM applications. This will grant a more complete picture of the uses of ADM. Similarly to some examples above, Angwin et al. (2016) discuss the COMPAS system. COMPAS is software that assesses recidivism risk of known offenders and is used in the US. The observed space of the algorithm consists of a set of features pulled from criminal records (Angwin et al., 2016). These features include relatively standard information, such as age or the number of times the offender has been arrested. The features also include information that is (at least) a bit questionable for the purposes, such as whether the offender's parents are separated or how often the offender felt bored in the past months (Angwin et al., 2016). The information from the observed space is used to assess people on probation.

Allhutter et al. (2020) discuss an ADM algorithm employed by the Public Employment Service Austria, that classifies Austrian job seekers into three categories, indicating their chances of finding a job within a certain time period. The features in the observed space include gender, age, highest level of education, and health impairment. Based on the algorithm's classification, job seekers get offered different support in (re)entering the labor market.

Another case can be found in Langenkamp et al. (2020), who discuss algorithms that can be used in the hiring process of companies or institutions. Such algorithms could for example predict the chances that a job applicant's contract will be extended after one year. Features in the observed space might include age, highest level of education, or the length of the previous employment. This information can advise employers in the decision to hire a job applicant or not. In fact, systems like these have been used by large employers, such as Amazon (Dastin, 2018).

## 2.2 The Fairness of Algorithmic Decision Making

Considering the examples in the previous section, it becomes clear that ADM systems and their applications concern and affect real people's lives. In fact, ADM systems are widespread and help make decisions that have far-reaching impacts on individuals and society as a whole (Ntoutsi et al., 2020). Recognition of this fact has raised the awareness of the importance of *fair AI*: ADM systems should not be biased against certain (groups of) people and should make fair decisions for everyone.

### 2.2.1 The problem

If we view ADM as a mapping from the construct space to the decision space, Binns (2018) states that "questions of fairness might arise if we suspect that distances between individuals in construct space and observed space differ depending on gender, race, religion, or other salient groupings" (p.74). Ideally, we would want the construct space and observed space to be identical (i.e., there is no distance between them). However, as this is often impossible, we strive for a minimal distance between the two. If the distance between construct space and observed space differs between individuals based on any of the 'salient groupings' Binns mentions, this might mean that the predictions are accurate or favorable for one group over the other.

For example, consider an ADM system that is used to assist in hiring decisions, where the construct space consists of actual job-relevant knowledge and the observed space consists of professional qualifications (Binns, 2018). If the distance between construct space and observed space differs by gender, men might require less job-relevant knowledge compared to equally qualified women. That is, because of several biases and unfair inequalities in society, it might generally be easier for men to obtain professional qualifications, compared to women. This means that women have to possess *more* job-relevant knowledge to achieve *the same* professional qualification. As a result, the observed space gives a reasonable representation of the construct space for men, but not for women. Women will likely be disadvantaged by the algorithm, which is not fair.

A deeper question remains whether the chosen features in the observed space are appropriate grounds on which a decision can be based. For example, is it legitimate to measure a person's job-relevant knowledge through their professional qualifications? Or, to return to the COMPAS algorithm (Angwin et al. (2016)), to measure recidivism risk through how often a person felt bored in the past months? Of course, ADM algorithms learn correlations between a large *collection* of features, not just on a few. However, algorithms cannot distinguish between less relevant features and more relevant features. The danger thus exists that less relevant features somehow come to weigh heavily on the prediction. This danger is exacerbated by the fact that many machine learning algorithms are opaque, which makes it very difficult (or impossible) to see the weight of features in a decision. I will argue that determining the features in the observed space is one of the major ways in which values enter ADM.

Another common problem for ADM is that data sets are not always representative of society as a whole. Especially with Big Data, it is often the case that some groups are under-represented, while other groups are over-represented. Misrepresentation in the data often results in vicious cycles of discrimination (Barocas & Selbst, 2016). For example, Buolamwini & Gebru (2018) found that facial recognition tools generally perform significantly better on lighter male subjects than on darker female subjects. Another example has been found in an algorithm that was applied by Amazon, to select top candidates for a particular job position. Dastin (2018) states that this algorithm is biased against women, as most data the algorithm was trained on is male (the tech industry being dominated by males). Because the algorithm had mostly seen males that were successful in certain jobs, it learned that being male is an appropriate precondition to success.

Lastly, Lepri et al. (2018) warn that ADM's lack of transparency is dangerous. Not only does opacity of systems pose a threat on their fairness, it also leads to a situation of information asymmetry, where a powerful few have resources and knowledge to access and control ADM systems that target a majority of society. This poses a threat on important democratic values, such as the importance of fair procedures.

### 2.2.2 Technical approaches to fair AI

In the field of fair AI, there is a wide range of technical approaches to create fairer AI. Technical approaches do not focus on the (societal) context in which a system is applied. Rather, they focus on understanding, mitigating and eliminating those internal aspects that make algorithms unfair. These approaches often focus on bias. Ntoutsi et al. (2020) provide a broad overview of approaches to handle bias in AI-driven decision-making. They distinguish three different categories that can be seen as different steps in the algorithmic process: (1) understanding bias, (2) mitigating bias, and (3) accounting for bias. I will briefly describe the categories below. However, it is beyond the scope of this thesis to give a complete overview of technical approaches to fair AI.

A widely adopted approach to understanding bias in AI is investigating how 'fairness' is defined. Because AI relies on computational operationalizations, a formal (i.e., mathematical, statistical) definition of fairness is required. Definitions of fairness are used to assess algorithms and evaluate how fair they are. The connection between bias and fairness is straightforward in this sense: if a fairness assessment reveals that an algorithm is not fair for, e.g., women, we say that the system is biased against women.

Verma & Rubin (2018) categorize existing definitions of fairness into separate groups: (a) predicted outcome, (b) predicted and actual outcome, (c) predicted probabilities and actual outcome, (d) similarity based, and (e) causal reasoning. Predicted outcome looks only at a model's outcome. For instance, 'demographic parity' looks at the percentage of minority and majority groups in the positive class (e.g., 'gets a loan', 'will re-commit a crime'). Predicted and actual outcome definitions compare the predicted outcomes with the ground truth labels. For example, 'equalized odds' dictates that the false positive and negative rates should be the equal among different groups.[3] Predicted probabilities and actual outcome looks at the predicted probabilities, instead of a predicted class. Similarity based definitions of fairness state that similar individuals should be treated similarly. Finally, causal reasoning definitions look at the relations between different features in the data set and their impact on the outcome of the system. For example, Kusner et al. (2017) introduce 'counterfactual fairness': the predicted outcome should be the same if a sensitive attribute of an individual is changed (e.g., a decision to not hire a woman for a job should be the same if only the gender of this person would be changed while the other features stay the same). Definitions that focus on causal reasoning try to ensure that ADM decision are not based on sensitive features.

While a formal definition of fairness is needed to evaluate the performance of AI systems, it is not at all obvious that such definitions are appropriate. Corbett-Davies et al. (2017) show the statistical limitations of several mathematical definitions of fairness. Furthermore, it is not immediately evident *which* statistical fairness definition should be applied in which context. To be safe, one could argue that it is best to satisfy all fairness definitions. However, according to Hedden (2021), this is impossible in practice. It thus remains unclear which definition to maintain. More generally, Ntoutsi et al. (2020) emphasize that, as the root of bias is not a solely technical problem, it is unlikely to find a fully technical solution to the problem. Different authors have suggested that fairness cannot be defined in purely mathematical formalism, because it is contextual, procedural, and dynamic (i.e., it changes over time) (D'Amour et al., 2020; Selbst et al., 2019). The sociotechnical status of fairness in AI is further highlighted by Mitchell et al. (2021), who state that "any definition of fairness necessarily encodes social goals in mathematical formalism" (p. 2). Ultimately, this means that any notion of fairness is meaningless outside of the particular context it is employed in. The formalization of fairness remains an open research problem in the field of fair AI.

On top of solely focusing on *understanding* bias in AI, the second approach to tackling bias in ADM systems is *mitigating* bias (Ntoutsi et al., 2020). Bias mitigation can occur in different stages of the machine learning pipeline. Preprocessing approaches focus on the data set, ensuring that the data set is as balanced as possible. The idea is that the fairer the data set is, the less discriminatory the algorithm will be. An example of a preprocessing approach is proposed by Kamiran & Calders (2009), who describe a method to change the ground truth labels of instances in the data set that are close to the decision boundary. In-processing approaches focus on changing the internals of a model to actively tackle biased behavior. This can be done by letting a fairness definition play a direct role in the training of the model, for example. As a result, the model's internal structure has built-in safeguards for fairness. Lastly, post-processing approaches focus on altering internals or outputs of a model, after it has learned from the data. For example, Kamiran et al. (2018) propose

[3] False positive rates refer to the proportion of instances that have been labeled as 'positive', while their ground truth is negative. False negative rates refer to the proportion of instances that have been labeled as 'negative', while their ground truth is positive.

a method equalize predictions among different groups by promoting or demoting predictions that are close to the decision boundary.

The third approach to tackling bias focuses on *accounting for* bias. According to Ntoutsi et al. (2020), algorithmic accountability refers to allocating the responsibility for how an algorithm is designed and what its consequences are for society. This can be done in multiple ways. A proactive approach is bias-aware data collection, for example by ensuring equal representation of different groups. A retroactive approach focuses on explaining algorithmic outcomes after the internals have been learned. *Explainable AI* refers to the goal of being able to explain the internal workings of a model in human terms (Ntoutsi et al., 2020). The idea is that if we understand why a model makes a certain (unfair) decision, it is easier to recognize this decision as biased, and do something about it. This explanation can be in terms of model *interpretability.* Interpretable AI focuses on explaining the inner workings of a system (Poursabzi-Sangdeh et al., 2021; Selvaraju et al., 2016), indicating what information in the data was used to make a prediction. Another example can be found in counterfactuals (Kusner et al., 2017). Counterfactuals explain if and how an outcome would change if *one* feature were different. For instance, one might see if a decision as to whether to grant someone a loan changes if one were to change the person's age, gender, race or zip code.

### 2.2.3 Contextual approaches to fair AI

Contrary to technical approaches to fair AI, contextual approaches focus on revealing and understanding ways in which the application of a system can be biased and unfair. Instead of making the internals of a model fairer, a contextual approach aims to provide a deeper understanding of a *why* a model is unfair. The intuition is that these insights can inform the technical interventions and the regulating policies that are needed to ensure a safe use of a model.

In order to understand bias, one could look at the sociotechnical causes of bias (Ntoutsi et al., 2020). Bias exists in society, often in the form of 'institutional bias'. Institutional bias is the tendency of procedures and practices of particular institutions (e.g., government, educational institutions, medical institutions, etc.) to operate in a way that disadvantages certain groups. The most common forms of institutional bias are racism and sexism. As ADM systems rely heavily on human data, it is inevitable that institutional biases are embedded in data sets. Looking at the complex sociotechnical systems that use these data sets (e.g., the Web), one can try to understand the ways in which such systems express or even amplify societal biases (Ntoutsi et al., 2020). It becomes evident how biased ADM systems may play a significant role in reproducing and amplifying pre-existing biases in society, uncovering the potential dangers of such systems. For instance, suppose an ADM system is used to decide whether a person gets a loan or not, and the system is biased against people of color. This is a case of institutional bias that can be traced back to many institutional processes in society. ADM will learn from biased data, only reinstating and reinforcing existing patterns, providing more 'evidence' for future biased decisions.

Apart from looking at *bias* exclusively, the way an ADM system is *applied* also reflects certain values. Thus, in order to examine how AI can become fairer, it is crucial to understand the context in which an algorithm is applied.

Recent examples of such thorough analysis can be found in Allhutter et al. (2020) and Crawford & Paglen (2021). First, Allhutter et al. (2020) examine an algorithm that is used to classify job seekers in categories based on their chances on the labor market. They explain how this system co-produces societal values. For instance, by predicting a person's chances to get a job as a function of their characteristics, reflects a individualistic world view. That is, a person's (in)ability to find a job is seen as a personal merit or defect, instead of a combination of effects concerning the individual *and* the labor market. Understanding how these societal values play a role in the algorithm, can help us identify the types of bias that might exist in the ADM system (e.g., there is a chance that a person's categorization is based on age, which is not desirable).

Second, Crawford & Paglen (2021) discuss the values that play a role in AI-based image recognition. Specifically, they focus on the practice of image annotation, a process that is crucial for supervised learning image recognition algorithms. Image annotation denotes the process of manually (i.e., done by a human) describing what an image is depicting. Crawford & Paglen (2021) show that the automatic interpretation of images is inherently social and political, because the manual labeling of images (especially images containing humans) depend on physiognomy: the assumption that people's personal traits are deducible from their looks. These assumption are largely shaped by institutional values and are often amplified by fictional works (e.g., villains are often portrayed as 'ugly'). Understanding these mechanisms, helps with understanding how automatic image interpretation can be biased, for example against people who do not conform to current beauty standards.

I believe that these thorough analyses of algorithms and the way they are shaped by values are fundamental for the development of fair AI, because it gives a more complete understanding of the biases that might be embedded in systems. The discussed works of Allhutter et al. (2020) and Crawford & Paglen (2021) provide a good example. However, these works are limited in scope, because they both focus on one specific applied algorithm. My aim in this thesis is to extend contextual analyses of algorithms beyond the scope of specific applications, and provide a taxonomy of values that might play a role in ADM in general.

A good example of a broader contextual approach can be found in Franzke et al. (2021), who propose the 'Data Ethics Decision Aid'. This tool accounts for bias by asking developers critical questions about the data and type of algorithm(s) used in ADM systems. This will make AI-workers more aware of the possible consequences of the choices they make during the design and developing stage, as well as informing policy-makers about the potential pitfalls and dangers of a system.

### 2.2.4 Feminist AI

There is a sub-field of AI research that deserves more attention here. The interdisciplinary field of *feminist AI* approaches AI from a feminist perspective. For instance, Adam (1995) criticizes symbolic AI (i.e., AI that is based on high-level symbolic representations, often in the form of logical expressions), by stating that it is based on a type of knowledge that is traditionally more ascribed to men. In this way, the objectivity and neutrality of (symbolic) AI is being questioned. The epistemology of symbolic AI is predicated on traditional rationalist epistemology, and Adam investigates this in two ways. First, she claims that the 'knower' or subject of knowledge is traditionally viewed as a white male (an 'expert'), and that the object of knowledge (i.e., what is known) is viewed as a propositional statement (i.e., observable knowledge). This view in itself shows a value commitment: it denies that knowledge is a cultural and social product and sees the subject as universal (i.e., a homogeneous group, often consisting of white males).

Adam claims that knowledge *is* a cultural product, and that not all knowledge is propositional. For instance, babies respond to people before they respond to physical objects, showing that primal knowledge is not propositional (Adam, 1995). Furthermore, the subject of knowledge is not universal, as different people may know different things. In the traditional epistemological view, a plurality of standpoints is disregarded as an inconvenience. Because the traditional bearer of knowledge is the white male expert, and the knowledge they produce is valued the highest, a hierarchy of knowledge is created, with women's knowledge near the bottom. Second, Adam (1995) discusses the difference between 'knowing that' (e.g., knowing that a bicycle is a means of transportation) and 'knowing how' (e.g., knowing how to ride a bicycle). Traditionally, knowing that is more ascribed to men, while knowing how is more associated with women. Symbolic AI is based on factual knowledge, or knowing that, and thereby disregarding knowledge that is valued by women.

Another example of feminist AI can be found in Wellner & Rothman (2020), who examine the possibility of creating algorithms that are not shaped by gender bias. They state that the common tendency to discuss bias in technical terms comes short, because AI involves complex relations between users, data sets, and algorithms. It is therefore crucial to understand these relations, and this requires more than just technical explanations. Understanding the relations between users, data sets and algorithms helps us map the solutions to gender bias in AI. Lastly, Wellner & Rothman (2020) use feminist theory to describe the roots of bias in AI. According to liberal feminism, the roots of gender bias lie in the data, not in the algorithm. Conversely, according to radical feminism, the root of gender bias lies in the algorithm itself, as it is shaped by existing gendered power relations in society.

One might find these approaches limited, because they only focus on gender bias (thus discarding others forms of bias, such as racial bias and ableism) or target only symbolic AI. However, I think the insights from feminist AI are valuable for two reasons. First, the arguments apply to other forms of bias and AI methods as well. Modern supervised machine learning is still mainly a form of 'knowing that' (e.g., knowing that a person has a high/low recidivism risk) and can be biased as a result of the people developing an algorithm or labeling data (Crawford & Paglen, 2021). Second, the arguments from feminist AI are useful, because they demonstrate the importance of understanding the complex historical and social relations between people, institutions and technology. This understanding is crucial for comprehending bias in ADM and finding solutions for it.

## 2.3 Scientific Objectivity

In the previous subsections, I have focused on ADM, its applications, and methods to increase the fairness of ADM systems. By now it should be clear that ADM, and AI in general, often behaves in a way that is biased, despite the claimed objectivity and neutrality of data sets, computers, and algorithms (e.g., "data do not lie", "computers have no values"). In order to investigate this supposed neutrality, I now turn to another field where objectivity is claimed and desired: science. The debate around objectivity in science is a central topic in the philosophy of science. In this section, this debate will be outlined.

### 2.3.1 What is scientific objectivity?

There are different definitions of scientific objectivity. A definition of objectivity should be both strong enough to be valuable, and flexible enough to be useful in practice. Reiss & Sprenger (2020) define three definitions of scientific objectivity that have dominated scientific research. First, *objectivity as fruitfulness to facts* characterizes objectivity in terms of accurately and factually describing the state of the world. In this sense, a claim is objective if it correctly describes some aspect of the world. Second, *objectivity as absence of personal biases* defines objectivity as freedom from personal preferences and experiences. It is generally thought that science is different from the arts and other human activities, because in science all personal opinions and biases are gradually filtered out and replaced by agreed upon evidence and methods.

Third, *objectivity as absence of normative commitments and the value-free ideal* describes objectivity in terms of value-neutrality. In particular, it is the *contextual values* that pose a threat on the objectivity of science. In contrast to *cognitive values* (e.g., simplicity, predictive power), that may play a legitimate role in some scientific processes, contextual values are more personal and political (e.g., justice, equality, pleasure). A scientist is not supposed to be inclined to accept a theory, just because it aligns with their personal values. The underlying idea is that there is a dichotomy between fact and value, so that one cannot and should not influence the other. Since the ultimate goal of science is thought to be the production of empirical knowledge, contextual values have no place in science.

The focus in this thesis will be on this last definition of scientific objectivity:

**Definition 1** (Scientific objectivity as value-neutrality)**.** Scientific objectivity as the absence of normative commitments and the value-free ideal.

Lacey (1999) defines three principal components of value-neutrality in science: impartiality, neutrality, and autonomy. Impartiality means that scientific theories should only be accepted or appraised because of the cognitive value of their contribution to the scientific discipline. For example, a theory can be accepted based on its explanatory power, but not on its potential to support a particular political agenda. Neutrality means that scientific statements make no normative claims about the world: they describe the world as it is, not as it should be. Autonomy means that the scientific agenda should be shaped by the desire to increase scientific knowledge, and not by any external factors. The latter can happen when research is biased towards interests from large and powerful external structures, such as the tobacco industry or large pharmaceutical firms. The three principal components of value-neutrality in science can be seen as requirements for objective science (Lacey, 1999).

### 2.3.2 Cognitive and contextual values

To better grasp the meaning of value-neutrality, a short explanation of *values* is appropriate. Douglas (2009) defines two categories of values: cognitive values and contextual values. The ideal of value-neutrality in science relies heavily on the dichotomy between cognitive and contextual values: generally only cognitive values have a place in science. Cognitive values are those “that help one think through the evidential and inferential aspects of one’s theories and data” (p. 93). For example, simplicity is a cognitive value, because simple theories are easier to understand and more straightforward to work with in practice; fruitfulness is a cognitive value because a theory that is fruitful and productive provides many new directions for future research. Letting a scientific decision be guided by such cognitive values does not impact the truth of statements and theories.

In contrast, contextual values reflect the personal values of a particular society or person. For example, privacy, justice, and freedom are important values in many Western countries, such as the Netherlands. In general, it is thought that contextual values have a risk of impacting the truth, thus leading to theories that are not entirely objective (i.e., factual, measurable). For instance, we tend to condemn scientific research that is biased towards the interests of large companies, for example in the tobacco industry, pharmaceutics, or Big Tech. In extreme cases, the intrusion of contextual values in the scientific process can lead to an oppressive and intolerant scientific agenda. For example, in Nazi Germany many important scientific works were discarded because its authors were Jewish. The infiltration of contextual values in science increase the risk that scientific theories are not faithful to the truth or that scientists are not open to the best explanation possible.

The underlying core concept of value-neutrality in science, is the so-called 'value-free ideal' (Lacey, 1999):

**Definition 2** (Value-free ideal)**.** The value-free ideal states that scientists should aim to minimize the influence of contextual values on scientific reasoning.

The value-free ideal is underpinned by the value-neutrality thesis:

**Definition 3** (Value-neutrality thesis)**.** Scientists can gather and process scientific evidence in a way that is not biased by contextual values.

It follows that the value-free ideal can be attacked by undermining the value-neutrality thesis. If the value-neutrality thesis is rejected, one must either conclude that the notion of scientific objectivity as value-neutrality should be discarded, or one must change the definition of the value-free ideal.

### 2.3.3 Attainability of the value-free ideal

Following this pattern, the critique on the value-free ideal in science can be generally distinguished in two categories: critique on the attainability of the value-free ideal, and critique on its desirability. We will start with the arguments concerning attainability: is the value-free ideal realistically possible?

Lacey (2018) argues that the value-free ideal relies on a dichotomy between facts and values. That is, there is an inherent contrast between facts and values, making them mutually exclusive. After all, the value-free ideal prescribes that, in order to gain knowledge of objective facts, one should get rid of subjective values. This line of reasoning indicates that fact and values do not go together. However, Lacey argues that this dichotomy is not obvious. For instance, methodological choices and evaluation criteria for scientific theories contain value judgments. After all, many ethical rules apply when one wishes to conduct an experiment involving human subjects. Here we see that the distinction between fact and value becomes blurry. If such a dichotomy does not exist, it is not at all obvious that value-free science is even possible.

The lack of distinction between fact and value is also highlighted by Anderson (2004). She argues that value-neutrality bears upon two sub-arguments. First, the *psychological argument* states that scientists are biased and preoccupied if they act with values in mind. It should not be the case that scientist disregard 'inconvenient facts', solely because it does not line up with their personal or societal values. However, this argument presupposes that there can be empirical evidence for certain value judgments (after all, otherwise there could be no inconvenient facts). This observation directly contradicts the second sub-argument for the value-free ideal. The *logical argument* states that there is no deductively sound way to move from values to facts. Anderson (2004) denotes this with the statement 'science is value-free': the goal of science is to produce facts, and since values cannot play any role in this process, there is no place for values in science. However, if the dichotomy between science and values is true, then it should also be the case that ‘values are science-free’: there is no sound way from values to facts. That is, it should not be possible to use factual knowledge as evidence for certain values that we hold. Anderson shows that this is not the case, because we use past emotional experiences as appropriate evidence for value judgments. For instance, we tend to judge things that make us happy as good. Thus, the dichotomy fails: values are not science-free, from which it follows that science is not value-free. Consequently, the *can* be a place for values in science, contrary to the value-free ideal. Lastly, we can acknowledge that our value judgments can be mistaken. According to Anderson (2004), this proves that values and facts do not exist in separate spheres.

While the previously mentioned arguments attack the distinction between fact and value, another line of arguments focuses on the distinction between cognitive values and contextual values. If there is no clear distinction between cognitive and contextual values, we cannot hold that science is or should be value-free. After all, most would agree that there is *some* place for (cognitive) values in science, as we have seen in the case of ethical guidelines in human subject research. Crasnow (2004) highlights an important oversight of value-free science. In the value-free ideal, the ‘knower’ of science is seen as an individual. However, as is also emphasized by Longino (1996), the knower of science is not an individual, but a community of knowers. Because this community is diverse, it essentially contains multiple social values. This means that the cognitive values needed to understand scientific theories and processes are not necessarily accessible to everyone in the community in a way that is unmediated by contextual values. If this is the case, there is no clear distinction between contextual and cognitive values, as contextual values come before cognitive values. This claim is also endorsed by Douglas (2009), who states that epistemic (i.e., cognitive) values are always trained by non-epistemic (i.e., contextual) values.

### 2.3.4 Desirability of the value-free ideal

The second line of critique against the value-free ideal in science can be defined in terms of the desirability of the project: should we even *want* science to be value-free? Halpin (1989) argues that the value-free ideal causes science to enable oppressive social policy. The traditional goal of scientific objectivity demands that scientists separate themselves from the object of study, creating a separation between ‘the self’ (i.e., the scientist) versus ‘the other’ (i.e., the object of study).

Since traditionally most scientists were white, Christian, heterosexual middle-to upper-class males, this demographic has become the collective ‘self’. Consequently, everyone and everything that deviates from this group become ‘the other’, a deviant and inferior group. While the self is considered as valuable and normal, the other is in need of explanation, and often defined in terms of *how* they differentiate from the self (Perez, 2019). Halpin (1989) shows how the universally accepted conception of scientific objectivity provides a justification for the oppression and domination of all beings and things that are categorized as ‘the other’ (e.g., other genders, sexual orientations, ethnicities, religions, etc.). Let us take gendered oppression as an example. In the medical domain, we see a general lack of interest in the female body, which has existed throughout history (Perez, 2019). This has caused it being more difficult to detect diseases in female bodies, putting them more at risk than males. Furthermore, as we have seen with Adam’s feminist AI, women’s knowledge has systematically been downgraded and labeled as ‘spiritual’, ‘subjective’, or even ‘witchcraft’ (Adam, 1995). Not surprisingly, these terms are often used to denote the opposite of ‘objective’, which remains the golden standard.

In order to accommodate a notion of science that moves away from oppression, Longino (1995) proposes new (feminist) values that should have a place in science. Longino first defines the five constitutive values that *do* play a role in objective science, extracted from Kuhn (1977): accuracy, simplicity, internal and external consistency, breadth of scope, and fruitfulness. These cognitive values have a place in science because they have generally been thought to aid the scientific process. Adhering to these values increases the likelihood of a theory to be true. However, Longino (1995) claims that the dichotomy between contextual and cognitive values is false, as “social or practical interests function as so-called cognitive values in determining what counts as good or acceptable scientific judgment” (p. 383).

Longino (1995) denotes this view as ‘contextual empiricism’: the claim that, while experience provides the least defeasible justification for scientific statements, the evidential relevance of particular experiences is mediated by several background assumptions that operate on different levels. Longino claims that in certain theoretical contexts, the only reasons to prefer a traditional scientific value over a non-traditional value are of a socio-political nature. This can be seen as proof that the traditional values are not purely cognitive, which has also been demonstrated by Crasnow (2004) and Douglas (2009).

To account for the background assumptions that mediate science, Longino (1995) introduces new values for objective science. *Empirical adequacy* is concerned with the extent to which the observational claims of a theory match with the data. *Novelty* as a value can be seen as a way to question mainstream theoretical frameworks. *Complexity of interaction* recognizes that relations are complex and not unidirectional. *Ontological heterogeneity* treats a multitude of different entities as equals. While the previously mentioned values are mainly theoretical, the next are more practical. *Applicability to current human needs* favors research programs that generate applicable knowledge that is relevant to the problems society faces. Lastly, *diffusion of power* favors research programs that do not limit participation and utilization, for example by (unnecessarily) demanding expensive materials or an extremely high degree of expertise.

### 2.3.5 Roles of values in science

On a more structural level, Douglas (2009) states that the value-free ideal is inappropriate. She proposes an extensive overview of the structure of values in science, in what processes they can be found, and what roles they might play. The value-free ideal of science is not desirable, because scientists should consider the potential social and ethical consequences of the errors in their work, and they need social and ethical values to do so. However, it would be a mistake to disregard the value-free ideal altogether, because there must be some constraints as to when and how to apply contextual values in the scientific process. After all, a limitless and unstructured opportunity to use contextual values in science could, in very bad cases, lead to a situation similar to the example from Nazi Germany.

For this reason, Douglas (2009) provides an overview of the structure of values in science. She starts by making a clear distinction between the two *roles* that values might play in science: the direct and indirect role. To illustrate this distinction, consider the example of a scientist who needs to decide whether to accept or reject a theory based on the available evidence. If a scientist uses values in a direct role, the values act as reasons in itself to accept or deny the theory. For example, a scientist might accept a proof of global warming, not because they have substantial data to back it up, but because they are a climate activist.

On the other hand, if a scientist uses values in the indirect role, the values help determine what should count as sufficient evidence for the claim. For example, the proof of global warming is accepted because denying it could do serious harm to nature and people, and the value of preventing harm is bigger than the value of having fully conclusive evidence. According to Douglas (2009), in this stage of the scientific process (i.e., where scientific claims are accepted or denied), the indirect use of values is legitimate, while the direct use of values is not.

There are several places in the scientific process where there *is* a legitimate direct role for values in science, particularly in the early stages of a research project. When deciding what to research, values might play a direct role in shaping scientists' choices. Such choices are usually influenced by what the scientist finds important, what are pressing matters for society, what kind of research can receive funding from governments or corporations, and which research projects are ethically acceptable. For example, it is legitimate to investigate sources of renewable energy, because climate change is a major threat to society. Values can also be used in a direct manner when choosing the methodology of a research project. Especially when human subjects are involved, ethical values play a direct role in ensuring the subjects are treated with respect and that no harm is caused. Lastly, funding decisions in science embody the direct role of values as well. Funding decisions are usually based on what scientist and investors find important. For example, cancer research might be funded because the disease affects a lot of people, which makes it valuable for society (both healthcare-wise and money-wise).

Douglas (2009) emphasizes that the direct use of values should be limited to the early stages of the scientific process, where one decides what research to pursue and how the research should be carried out. In later stages of the scientific process, one can only use values in a direct way in special circumstances. For instance, a scientist may change the methodology in a later stage if the old methodology turns out to be unethical, but not because the old methodology does not produce the desired results. Similarly, a scientist should not accept or reject a theory solely because they do not like its implications. Note that the limited role for direct values does not only apply to contextual values, but also to cognitive values: while simplicity might be a legitimate cognitive value for science, a theory should not be accepted only because it is simple.

While we have seen that a direct role for values can, in some cases, undermine the goal of science (to obtain truth), this is not the case for the indirect role for values. This indirect role concerns the sufficiency of evidence, weighing of uncertainty, and the consequences of error (Douglas, 2009). For instance, if there is substantial uncertainty about a claim, and the consequences of error are known, values should determine whether the evidence is sufficient to support the claim. In this way, values are used to weigh the importance of uncertainty, but not the claim itself. Douglas (2009) not only states that the indirect role of values is *allowed* in science, but also that it is *needed*: "as long as science is inductively open, uncertainty is ineliminable, and thus so are values" (p. 114).

Douglas goes on to explain that the desire for value-free science stems from a confusion between value-laden science and politicized science. As we have seen, there is a legitimate role for values in science, because values are needed to judge the importance of errors, and to choose what research to pursue. However, politicized science happens when the direct role of values is not limited to special circumstances. For example, a scientist might conclude that smoking is healthy because their research was funded by a tobacco company. Politicized research often undermines the truth, while value-laden research does not necessarily do so.

As a last remark, it is important to note that a direct role for values does not necessarily mean that values *consciously* enter the scientific process. For example, a scientist may not be aware that they are inclined to accept a theory because they value environmental justice. Nonetheless, most of us would still argue that the direct role for values in this stage of the scientific process (i.e., accepting or rejecting a scientific claim) is not desirable: even if one values environmental justice, the direct role for values in science should stay limited in order to protect the objectivity of science. Similarly, when values play an indirect role, their use is not necessarily *unconscious*. For example, a scientist often consciously uses values to determine the importance of the potential consequences of their errors. If a scientist chooses to accept a theory that confirms climate change despite not having definite proof, they can do so because they have consciously used values to reason about the social importance of the topic.

In this section, I have outlined the debate around value-neutrality in science. It should be clear now that complete absence of values in the scientific process often is a myth, because facts and values do not operate on different spheres. Moreover, in different parts of the scientific process, values *can* and *should* play a significant role, for example to make sure that experiments are ethical. Thus, the strict value-free ideal fails and should at least be reformulated into a more permissive statement. Now that we can conclude that science is not entirely value-free, let us turn to the status of values in ADM systems. Do values play a role in ADM? How do values enter in ADM? In what steps of the ADM process can values be found, and what kind of values are these? I will answer these questions in the next section.

# 3 Value-Neutrality in Algorithmic Decision Making

In this section, I will answer the first and second research question of this thesis:

1. How does the value-neutrality of science relate to the claimed objectivity of algorithmic decision making?
2. To what extent can the philosophical arguments *against* value-neutrality of science be applied in the context of the objectivity of algorithmic decision making?

I will argue that value-free ADM is not possible and not desirable. In some cases this is problematic, but is does not have to be. This section will first describe the parallels between science and ADM, focusing on similarities in epistemology and the unconscious use of values. Second, the discrepancies between science and ADM are illustrated. These differences lie in the distinctions between cognitive and contextual values in both fields, and the different legitimacy of the direct role for values.

## 3.1 Parallels Between Science and ADM

Ultimately, my aim is to show that ADM, like science, is not value-neutral. In this section, I will describe the similarities between both fields. I will identify the arguments against the value-neutrality of science, that can be used to expose the value-ladenness of ADM as well. I will argue that the first similarity is based on a shared epistemology, while the second similarity lies in the way in which values unconsciously influence processes in both fields.

### 3.1.1 Epistemology

First, I will turn to the epistemology of science and ADM. Epistemology refers to a branch of philosophy that is also denoted as the theory of knowledge. In particular, I will look at humans' quest for objectivity, which is represented in both science and ADM. More importantly, I will draw a similarity in *whose* knowledge is represented.

The first similarity with regard to epistemology, can be found in humans' quest for objectivity. This similarity points to the belief that value-free knowledge is valuable and attainable, as well as to the methods that are thought necessary to obtain value-free knowledge.

Both the endeavor of science and ADM are respected because we believe that objectivity is important and that both fields are value-neutral. When it comes to ADM systems, the general line of thought seems to be that their predictions or classifications are objective and value-free (McQuillan, 2018). This point is reinforced by Lepri et al. (2018), who state that the shift towards ADM and Big Data can be described as satisfying a demand for greater objectivity. This 'hunger' for objectivity can also be found in science. According to the value-neutral ideal of science, the ultimate goal is to gain objective insights into how the world works (Reiss & Sprenger, 2020).

Moreover, this demand for greater objectivity is associated with the value-neutrality thesis. That is, there is a belief that obtaining objective, value-free insights, is possible. After all, if we were to think that value-free knowledge is impossible to achieve, then why would we make value-freeness one of the main pillars of science (or even bother to do scientific research at all)?

It is thought that a data-driven approach can provide access to objective truths. Both science and ADM use data to calibrate their theories or systems. For instance, a scientist carries out an experiment to obtain data that can be used to evaluate hypotheses; ADM systems use ground truth data points to tweak weights in the model, to ultimately produce the their optimized decisions.[4] Furthermore, both science and ADM make inductive leaps from past to future data. Scientists use past data, obtained from observations and/or experiments, to induce theories that can (help to) predict or explain future events or behavior. These theories can in turn be tested on new data. In a more direct way, ADM often *re*produces patterns that were found in past data sets (Allhutter et al., 2020).

The second, and arguably more important similarity between science and ADM, is concerned with their 'bearer of knowledge'. In other words, who is the 'knower' of science/ADM? Whose knowledge is represented? In general, it is believed that the value-free aims of science and ADM go together with a value-free knower, a 'view from nowhere'. As a result, the individual knower of science and ADM is unimportant and the existence of a 'universal knower' is assumed.

As we have seen, Longino (1995) criticizes the common idea of the 'universal knower' of science. First, she argues that 'the view from nowhere' is not attainable, because humans approach things from their individual perspective. Longino argues that science consists of a community of knowers (i.e., different scientists), where each individual brings along certain social values. Second, because scientist have historically mostly been white males, the 'universal knower' of science is not universal at all. It is the knowledge of this demographic group that is represented in science, not so much the knowledge of others. They do not have a 'view from nowhere', but rather a view that was shaped by particular experiences which helped shape them.

Turning towards ADM, we have seen a similar argument from Adam (1995): the endeavor of creating machine intelligence is based on factual data and reflects a prioritization of knowledge that is typically assigned to male characteristics. This insight indicates that the knowledge on which ADM is based, does not stem from a 'view from nowhere', but rather from a particular societal perspective. Like in science, of the power within ADM lies with white, middle-aged males. Thus, ADM systems are based on the knowledge, (value) judgments and perspectives of this collective group.

Both Longino (1995) and Adam (1995) argue that a pluralistic concept of the knower would help science and/or AI forward. This concept should take into account multiple different perspectives and (value) judgments. Because a pluralistic concept of the knower is often still lacking, both science and ADM can enable a dangerous hierarchy between 'the self' and 'the other'.

Halpin (1989) argues that the value-neutral ideal in science, together with the dichotomy between emotions and the intellect, provides a justification for the oppression of 'the other'. The value-free ideal encourages a separation between the scientist ('self') and the object of study ('other'). Likewise, the rapid growth and potential of ADM creates a knowledge gap and power imbalance between developers, users (i.e., governments, institutions, private companies or individuals who employ ADM in their procedures), and people who experience the consequences of ADM in practice (i.e., the people that ADM systems make decisions about). Thus, through ADM, developers and users ultimately make decisions about individuals that do not have access to resources and tools that are needed to understand what is happening (let alone being able to object to a decision).

[4] This may depend on the goal: sometimes a most accurate prediction may be desirable (e.g., closest to the truth), while other times a fair distribution of outcomes is more important.

### 3.1.2 Unconscious use of values and positionality

The second parallel between science and ADM concerns the way in which values enter the process unconsciously. The value-free ideal is present in both science and ADM: both fields aim to obtain value-free, objective truths and try to do so with value-free methods. However, in both fields we will see that processes get unconsciously influenced by value judgments, which makes one question the value-freeness of science and ADM.

First, it can be dangerous when values unknowingly enter the process in science and ADM. In science, this can happen when scientists overlook certain facts or demographics. For example, in the medical sciences, menstruation and its impact on people's lives has largely been neglected. This is presumably the case because most (successful) scientists in the field did not have to deal with menstruation themselves (Bobel et al., 2020). Moreover, the actual content of science can be biased too. For example, research that focuses on optimizing car safety often uses dummies that are based on average male body measurements. This has caused cars to be less safe for people that do not conform to these measurements (Perez, 2019).

In ADM, gaps of expertise of developers can lead to situations where a system might not work equally well for all demographics. For example, it was found that AI software in cameras, that was designed to warn photographers when someone is blinking, seems to think that Asian people are constantly blinking (Zou & Schiebinger, 2018). Furthermore, similar topics to those neglected in science, tend to be overlooked in AI. For instance, when Apple Health launched in 2014, users were promised to be able to track a wide variety of important and interesting health metrics. Unfortunately, it did not include a period tracker (Perez, 2019).

Moreover, even the fact *that* ADM systems do not behave equally fair for all demographics is not always clear to developers and policy makers. For instance, Angwin et al. (2016) demonstrated that the COMPAS system operated in a way that is more advantageous for white people than black people, even though the accuracy of both groups is similar. It turned out that the number of false positives (i.e., offenders that are falsely predicted to re-offend) in the pool of black offenders was much higher than that of white offenders. The failure to recognize this unfair trait of the COMPAS system, indicates a shortcoming of its developers and employers.

All in all, in both science and ADM we see a biased standpoint of its initiators (i.e., scientists, sponsors, developers, deployers), which leads to theories and systems that are not equally beneficial for all people. Even more, there are research topics that are highly relevant for a large number of people that have been neglected because the majority of people in charge do not share the same experiences and needs. The biased positionality of scientists and ADM workers also makes it more difficult for them to recognize *when* a theory or system is harmful for certain groups, or when they disregard particular topics that are relevant for others. We see an argument that we have encountered before: the 'knower' does not have an unbiased, value-free perspective ('view from nowhere'), but approaches topics, methods, and evaluations from their own value-laden perspective.

The individual positionality of scientists and ADM experts cause them to unconsciously use values in processes/research aims that are supposed to be value-free. The tendency to favor things or people that correspond to one's own positionality is referred to as 'in-group bias'. This is not necessarily a sign of weakness or malevolence, but a general aspect (or limitation) of human cognition (Hedden, 2021). Even if people are aware *that* their perspective is not value-free, it is often nearly impossible to fully grasp *how* it differs from other positionalities. This is partly caused by another type of cognitive bias, called 'projection bias'. Projection bias refers to the fact that people tend to assume that their own way of thinking about things is typical or universal for all people (Perez, 2019). The cognitive limitations of humans make me believe that value-free ADM, like value-free science, is not possible to create.

Second, the widespread faith in value-free science and ADM reflects certain values in itself. These values often go unnoticed and unspoken. It is because of a societal value that we think science is a valuable pursuit (Douglas, 2009). Because of this existing value in society, we want to have a more complete understanding of the world and we think that science is the way to achieve this. This could be different, for example if our society would value stability above truth and wants to avoid the disruptive changes that science can cause.

Likewise, the sole pursuit of ADM intersects with certain societal values, in particular with specific forms of trust, efficiency, and objectivity (Rieder & Simon, 2016). ADM (and Big Data) could blossom in a society and political landscape that is characterized by strong feelings of distrust and uncertainty. This culture explains the need for greater objectivity. ADM is used as a pre-emptive tool to combat crime, treating all individuals as suspects. Furthermore, ADM's application in delicate processes is often justified because of its efficiency, both in terms of money and time. It is probable that ADM would not be as widespread as it is today if different values were in place (e.g., if the trust in the individuals were bigger, or if efficiency were less important than it currently is). Thus, while ADM claims to produce objective truths, its use in itself is justified by certain contextual values.

In sum, we have seen the parallels between science and ADM when it comes to epistemology and the demand for greater objectivity. Both science and ADM can be seen as a means to achieve the much needed objectivity that we, as a society, crave. Additionally, both science and ADM are in the hands of only a few who have access to the resources to fully grasp what is happening. This can lead to dangerous situations in which people who do not have these resources are oppressed. Furthermore, I have discussed the dangers of the unconscious use of values in both science and ADM. The positionality of scientists and ADM experts causes them to create models, procedures and systems that do not work equally well for all groups. Moreover, the fact that science and ADM are used and appreciated so much (as a means to achieve objectivity), is a value judgment in itself.

## 3.2 The Bigger Role for Values in ADM

Now that we have seen the parallels between science and AMD, let us turn to the differences between them. The differences that are mentioned in this subsection indicate that values might even play a bigger role in ADM than in science. The first difference concerns the *type* of values that may legitimately play a role in processes. Particularly, the distinction between cognitive and contextual values is less clear in ADM. The second difference concerns the *role* that values can legitimately play. I will argue that it is justified to use values in a direct manner in ADM, that would never be allowed in science.

### 3.2.1 The bigger role for contextual values

We start by focusing on the distinction between cognitive and contextual values in ADM. In science, we have seen that the dichotomy between cognitive and contextual values fails, because cognitive values are always trained by contextual values (Douglas, 2009). That is, cognitive values do not exist in isolation, but are a product of society, power structures, and culture. The same argument can be made for ADM, and I will argue here that the distinction between contextual and cognitive values is even less clear in ADM. This is because cognitive values are less well defined in ADM.

Recall that cognitive values in science are the values that scientists can legitimately use to think about the inferential impact of evidence/data, the risk of making errors, or the impact of their research (Douglas, 2009). Such values do not impact the truth of scientific theories For example, explanatory power is a cognitive value because a theory with large explanatory power potentially has more impact in practice (e.g., because it can be applied in different contexts), and thus may legitimately be used to prefer one theory over another.

In the context of ADM, we might say that cognitive values are those values that legitimately guide developers when thinking through the design and consequences of their systems. For example, simplicity might be a cognitive value because decisions of simple models are easier to understand and explain than decisions of complex models; accuracy is a cognitive value because accurate models give more adequate decisions that we can base further action on.

However, there is an important difference between science and ADM, which causes differences in what values can legitimately be used to think through the inferential aspects of a theory or model. ADM is inherently much more intertwined with its applications, compared to science. While science can have different purposes (e.g., understanding, explaining, or controlling), ADM has only one: making predictions from past data.[5] That is, ADM is developed *because* it was meant to be applied in practice. Although applicability is important for science as well, for example to obtain funding, science's main goal *can* exist regardless of its application in everyday life.

Even more, ADM has a more direct influence over people's lives. Although science is also used in ways that affect people, for example to guide laws and policies, it is less likely that scientific theories will play a direct role in decisions *about* people that have far-reaching consequences (e.g., decisions about hiring). Because ADM has a more direct and far-reaching influence over people's life, it is also legitimate to use *contextual* values to think through the societal impact of ADM. For example, what impact does a system have on (in)equality? Or how can a system influence the power dynamic between governments and individuals? Judgments with respect to these kinds of questions are not only made with the help of Douglas' 'traditional' cognitive values, but also need contextual values like equality, inclusion, and justice.

Although scientists can, in a similar vein, write about contextual considerations of their work, they are not allowed to *act* on it. For example, in political science one would not want to let contextual values play a role in thinking through a theory about polarization, rejecting or accepting a theory because is is not fair. Conversely, in ADM, it is legitimate to weigh 'traditional' cognitive values against contextual values such as equality and equity. For example, in the COMPAS case, Angwin et al. (2016) concluded that equally distributed false positive/negative rates are more important than overall accuracy of the system. Because science's main goal is to obtain value-free truths (which can then be used to serve different needs), the weighing of cognitive values against contextual values will tend to result in an overall favoring of cognitive values. This is not the case for ADM, where we want contextual values to play a role to increase the applicability of a system.

In sum, in science, the cognitive values that are used to think through the evidential and inferential aspects of one's theory and data, serve a purely *cognitive* role: acquiring knowledge and understanding of the world. However, values that are used to think through processes in ADM need not be purely cognitive. As we have seen, sometimes a contextual value like fairness is a legitimate justification for making certain decisions. This is not to say that cognitive values are not taken into

[5] Note that making predictions is also *one of the* main purposes of science.

consideration in ADM. The traditional cognitive values, such as explanatory power and accuracy (i.e., how well a model performs based on test data for which the ground truth is available) remain core factors on which ADM is evaluated.

Whereas cognitive values in science are usually used to justify decisions based on 'objectivity', in ADM it is much more common to justify decisions based on other, contextual values. For instance, decisions can be based on how *fair* they make a system. Similarly, systems can be rejected because they are unfair. We would not see this in science, as scientific claims or theories do not have to be anything other than objective. The next subsection discusses a related point concerning the direct use of values. Because science and ADM sometimes seem to achieve different goals, it is much more accepted to use a direct role of values in ADM.

### 3.2.2 The bigger role for a direct role for values

A direct role for values in ADM is more accepted than it is in science. Recall that Douglas (2009) describes a direct role for values in science as values that act in a way that evidence normally would. This can occur when values are used as reasons in itself to deny or accept a certain theory or claim. A direct role for values in this stage of the scientific process poses a threat on the integrity of scientific research, and should not be tolerated (Douglas, 2009). However, a direct role for values in this stage of the process *is* permitted (and even encouraged) in the context of ADM. Suppose we could choose between two ADM systems, one being more accurate and the other being fairer. In this scenario, it is perfectly fine to accept/use the system that is fairer, *solely* because fairness is more important to the developer/user than accuracy. Even if the other system would have better performance scores, the choice for the fairer system can be justified.

The direct influence of values in ADM can go even further. For example, it is justified to adjust data sets in order to make ADM more/less fair or inclusive. A popular method to increase equal distributions of groups in data sets involves manually inserting data points of underrepresented groups (Kamiran & Calders, 2009). Thus, here we see that it is legitimate in ADM to *actively* let contextual values influence processes and methods. Note that this can be justified either because the data do not represent the world as it is (e.g., the data was collected from a source that is not representative of society as a whole), or because the data *do* represent the (unfair) world as it is and we want to do something about it (e.g., a group is underrepresented in the data because of institutional biases in society). In the first case, one makes a cognitive claim about the accuracy of the data. In the second case, one makes a normative claim about how the world *should* be.

Especially this last case does not align with the value-free ideal, which states that ADM experts should aim to minimize the influence of contextual values on scientific reasoning. Thus, we see an important discrepancy between science and ADM. The value-free ideal is not as clear-cut in ADM as it is in science. While it is generally accepted that the main purpose of science is to obtain knowledge and that the value-free ideal will help achieve this goal, ADM needs to be more than objective. It needs to be fair, so that we can use it in practice. However, since fairness is a contextual value, we *need* contextual values to guide ADM, thereby contradicting the value-free ideal.

The fact that we need contextual values to guide ADM is emphasized by the many interventions that are undertaken when "ADM goes wrong". For example, Langenkamp et al. (2020) describe approaches to make hiring algorithms fairer. They mention a hiring algorithm that was implemented by Amazon, and failed to classify women as suitable candidates. This algorithm was quickly put out of use, indicating that unfair ADM is not desirable. Second, Angwin et al. (2016) describe the

COMPAS system, that proved to exhibit racial bias. After this became known, a Supreme Court ruling determined that the tool's limitations and cautions should be made clearer. Interventions like these show that ADM needs to adjust to the context it is applied to for it to be useful. We do this by letting values like fairness play a direct role throughout the designing pipeline of ADM.

In sum, this section has focused on two main differences between science and ADM. First, I have demonstrated that in ADM, it is not only acceptable to use the 'traditional' cognitive values to think through the evidential and inferential aspects of a model, but it is also justified to use contextual values. Furthermore, it is legitimate to *act* on contextual values in the context of ADM, by letting them play a direct role. By allowing contextual values to play a major role in decision making in ADM, the value-free ideal of ADM is being questioned. The differences between science and ADM can be explained by the fact that their main goals are not identical. Whereas science is focused on obtaining objective truths and understanding, ADM aims at making *'good'* predictions. To serve this aim, ADM needs to be guided by cognitive values like accuracy *and* contextual values like fairness.

# 4 Values, Fairness, and Wrongful Discrimination

In the previous section, I have argued that ADM is not, and cannot be, value-free. We have seen that this can be dangerous, for example when ADM experts take their positionality to be universal and thereby exclude people that do not share their experiences. On the other hand, I have also argued that ADM is value-laden (i.e., not value-free) *because* we want contextual values, like fairness, to play a direct role. In this section, I will take a step back and see what these conclusions mean for (fair) ADM. I will argue that, while ADM is not value-free, this does not necessarily mean that systems only make unfair decisions. Since the previous section has demonstrated that value judgments are deeply intertwined with ADM systems, I will argue that it is more effective to *manage* values, instead of *omitting* them. In this section I try to stay descriptive, without making normative claims about what is good.

## 4.1 Different ideas about (un)fairness

To grasp the relationship between values and unfairness, it is important to take a step back and understand exactly *what* it is that makes a decision unfair. How does unfairness relate to (wrongful) discrimination? And what is its connection to biases in ADM? As we will see in this section, there are different schools of thought about what makes something unfair. The particular school of thought that is adhered to, will inform the set of values used to guide a decision.

Let me start by making a comment on what unfairness is *not*: unfairness is not the same as wrongful discrimination. That is, a decision can be unfair, without it being a case of wrongful discrimination. For example, if an ADM system decides who to hire for a job purely based on random choice, this does not seem to be fair. However, if every individual is affected in the same way by such a system, it is also not a case of wrongful discrimination. From this example, we can conclude that unfairness is not a necessary condition for wrongful discrimination. However, it does seem to be the case that wrongful discrimination is always unfair.

The question then remains: What makes an ADM-aided decision wrongfully discriminatory? Before we answer this question, it is helpful to look at the term *discrimination*. As Moreau (2020) notes, to discriminate means to differentiate between people or groups. Surely, not all cases of differentiation lead to wrongful discrimination. We differentiate between age-groups when deciding which individuals are allowed to purchase alcohol, and this does not seem to be unfair. The question of wrongful discrimination thus can be framed into a question of differentiation: when are differentiations unfair, and when not?

In the context of ADM, **luck egalitarianism** has a take on when differentiation is unfair. Luck egalitarianism prescribes that, while inequalities that are the result from one's own choices may be justified, inequalities that are due to pure luck are wrong (Binns, 2018). Thus, one may legitimately be treated differently because of their choice of clothing, but not because of their natural hair color. According to luck egalitarianism, COMPAS is wrongfully discriminatory because it bases its decisions on features of individuals that are due to pure luck, such as family circumstances. The inequalities that are a result of personal choices do not have to be accounted for. This leads us to the following definition of unfair ADM:

**Definition 4** (Unfair ADM according to luck egalitarianism)**.** ADM is unfair if different individuals or groups are treated differently and if that unequal treatment is based on aspects that are the result of pure luck.

Critics argue that sometimes it is necessary to compensate inequalities that are the result of personal choices, for example when individuals voluntarily place themselves in unequal positions to serve an important social purpose. Another objection to luck egalitarianism, is that in some cases it is permitted and even desirable to take into account features that are due to pure luck. For example, it is justified that student admissions councils take into account applicants' intelligence. Furthermore, luck egalitarianism does not account for the fact that basing decisions on random choice does not seem fair to us. According to luck egalitarianism, a decision may be based on random factors, as long as they are not due to pure luck. Thus, a decision about admitting an applicant may legitimately be based on the color of clothes they are wearing, according to luck egalitarianism. This does not seem desirable, and leads us to an adjusted definition of unfair ADM:

**Definition 5** (Unfair ADM according to relevance)**.** ADM is unfair if different individuals or groups are treated differently based on factors that are not relevant for the purpose.

Somehow, we only want to include *relevant* features in a model.[6] Ideally, these features would not be based on pure luck, but sometimes it is inevitable that they are (e.g., taking intelligence into account in the college application case). Admittedly, this definition of unfair ADM is rather banal. In my view, this banality points to the fact that it is not useful to define a universal definition of unfair ADM. Furthermore, it is not possible to create a general list of relevant factors, as they differ per application. Thus, one should look at the specific context in which a system is applied, to determine whether or not its decisions are unfair.

One last remark about Definition 5, is that it is not a fully conclusive definition. It could be the case that, while ADM is based on only relevant factors, its decisions still turn out to be unfair. For example, suppose a college only takes into account applicants' intelligence and their extracurricular activities, to decide whether or not to admit them. One could argue that these factors are relevant: they are good measures to predict someone's success rate in college. However, one could also argue that this system unfairly disadvantages applicants from poorer socioeconomic backgrounds, by not acknowledging that socioeconomic background plays a significant role in factors like these. For example, applicants from low-income families might be more likely to take a part-time job after school, preventing them from doing extracurricular activities for extra credit. This example again shows the importance of a context-based approach to tackle unfair ADM: whether something is unfair or not, is highly dependent on the context.

Our ideas about unfairness, based on luck egalitarianism and relevance, suit a broader societal value of individualism. It fits the great promise of the American Dream: everyone is equal and deserves equal opportunities. If one works hard, everything is possible, regardless of social class or circumstances at birth. All people are born with a free will, which enables them to make choices that make them better or worse off than others. With free will comes a responsibility as well: one must be able to bear the inequalities that can result from their choices. However, where inequalities are the result of circumstances outside an individual's control, they should be compensated. Examples are inequalities due to one's gender, race, nationality, family, socioeconomic background, etc. These factors are called 'sensitive features'. Including sensitive features in ADM will significantly increase the chance of wrongful discrimination.

[6] Note that we need something more than just *statistical relevance*, as values in society might cause statistical correlations between certain factors that are undesirable. For example, there may be a statistically significant correlation between gender and intelligence that is caused by the way intelligence is measured. However, we would not want gender to play a direct role in college admittance. Thus, features should be relevant on a more fundamental level, such that their use is explainable using common sense.

Thus, we generally should aim to avoid sensitive features in ADM systems. The problem is that inequalities are intersectional, meaning that individual features interact with each other in many complex ways. This makes it almost impossible to find a feature that is *not* related to any of the sensitive features. For example, a poor socioeconomic background may motivate someone to drop out of high school. While their choice to drop out may not be 'due pure luck', their socioeconomic background is. To complicate the matter, ADM systems do need *some* personal information to base decisions on.

When evaluating the fairness of ADM systems, it helps to keep the definition of luck egalitarianism in mind and understand its criticisms. Recognizing that this particular definition of unfairness is itself also based on certain societal values and making these values explicit, may help gaining insight into where the system can go wrong. For instance, the Austrian ADM system for predicting an individual's chances on the labor market (Allhutter et al., 2020) might be discriminatory because it uses features that are based on pure luck (e.g., gender). Other features, such as health condition, are perhaps more likely to be the result of personal choices. However, poorer health conditions are often related to a poorer socioeconomic background, which in turn is a feature based on luck. Thus, letting health condition play a role in determining labor market chances might be problematic as well. Understanding these mechanisms allows one to see a more complete picture of an ADM system and recognize its potential pitfalls.

Now that we have a reasonable definition of unfair ADM and we have seen its complications, let us turn to our main topic: values in ADM. What role does the value-ladenness of ADM play in the debate around what is fair? A system is unfair *because* it is biased, but the Amazon hiring example shows that unfair ADM is not necessarily caused by a direct role of values. The system is based on an 'objective' data set (in the sense that it is representative and stays true to the observed truth), but draws the wrong conclusions from it.[7] However, according to our definition of unfair ADM, the system *is* wrongfully discriminatory, because women are treated differently purely based on a factor (i.e., gender) that is the result of pure luck. One could also argue that the system is unfair, because gender is not relevant when looking at someone's potential in Tech. Here we see the importance of the 'fundamental' level of relevance: while gender indeed *in theory* should not matter, the data somehow tells us that it *does*. This example demonstrates how difficult it is to pinpoint exactly what it is that makes a decision unfair.

The Amazon hiring example shows an important insight: even if an ADM expert could act without values, this would not ensure that ADM is fair. The concept of value-freedom, as we know it from science, focuses on creating a distance between facts and social, emotional and personal aspects that might influence a process. As a result, data that is used in ADM is regarded 'objective' in the same sense as scientific data is. We have seen arguments that contradict the objectivity of scientific data, for example because the people who collect data can not approach them from a 'view from nowhere'. The same can be said to contradict the objectivity of data in ADM. Furthermore, data that is used in ADM is inherently social and intertwined with contextual values. This misinterpretation of data can fuel the idea of the objectiveness or neutrality of ADM.

Next, can we go a step further and claim that a value-laden ADM system can be fair? In previous sections, I have briefly touched on the topic of manually inserting values to ensure the bigger goal of fairness. This happens with a number of techniques that are used in fair AI. To shine more light

[7] Of course, values do play an indirect role in the sense that the available data is skewed because of certain values that have existed in society for a long time (preventing women from becoming Tech experts).

on this topic, the next subsection will discuss ways in which values can be used “for the better”, to make ADM fairer.

## 4.2 Using values for the better

While I have stressed that the (unconscious) use of values in ADM systems is dangerous and can lead to unfair discrimination, it is important to note that not all values are bad. There are important values in society that safeguard people’s rights. Examples are justice, liberty, equality, privacy, freedom, and fairness. We have already seen that decisions in ADM may legitimately be guided by such values in a direct manner. Current methods in fair AI often use these values to manipulate ADM systems.

For example, there are methods to mitigate bias by inserting artificial data points in a data set to make it less skewed (Ntoutsi et al., 2020). This way, more female data points could be added to the data set used for the Amazon hiring algorithm, mitigating the gender bias that the system used to show. Thus, this method to mitigate bias in ADM can be seen as a way to ensure fair representation in a data set.[8] In other words, values such as fairness and equality can consciously and directly be used to improve ADM systems. ‘Good’ values (that are dependent on the context) are manually inserted to make the system better.

Another example of the insertion of ‘good’ values in ADM systems can be found in the pursuit of designing explainable AI. It is generally thought that making algorithms more explainable is beneficial, because it makes it easier to trace back the algorithm’s reasoning steps. In turn, this makes it easier to detect the system’s reasoning mistakes that might result in unfair outcomes. Thus, we see that the value of fairness plays a direct role in this method. However, explainable AI also reflects a need for transparency, accountability, privacy, and justice. For instance, explainable algorithms help ensure a fair procedure (i.e., people have the right to explanation when something is decided for or about them), which is an important aspect of legal justice.[9] Additionally, making ADM explainable can also help increase its acceptance in society. Thus, different kinds of ‘good’ values can be used in the effort of making AI fairer.

Now that we have seen that current approaches in fair AI use values for the better, can we take it one step further and introduce new values that should guide ADM? Recall that something similar has been done for science by Longino (1995). Longino proposes new (feminist) values that should have a place in science: empirical adequacy, novelty, ontological heterogeneity, complexity of interaction, applicability to human needs, and decentralization of power. To what extent can these new values for science be used in the context of ADM?

It appears to me that these values either do not transfer well to ADM, or are already (albeit implicitly) used in fair AI methods. Values that should not be core values in ADM are empirical adequacy and novelty. While empirical adequacy is important for ADM to some extent (we would not want systems to be completely random), there lies a danger in holding onto this value too much. The Amazon hiring algorithm clearly demonstrates this point: while the data set that was used was consistent with empirical facts about the current situation, the system produced outcomes that were not desirable. Novelty as a value lies on the other end of the spectrum, prescribing that theories or systems differ in significant ways from mainstream theories/methods. This does not help ADM, because it is vital that systems are based on actual data, even if a system is designed to

[8] Note that the representation in this case *is* accurate, but does not lead to fair outcomes.

[9] GDPR Article 22, Recital 71.

find new patterns. Past data are the core building block of ADM. Furthermore, if ADM were based on data that does not conform to the world as it is, there lies danger in the fact that decisions are largely based on random choice. As we have seen, this is not desirable or fair.

By contrast, the other values proposed by Longino (1995) can be useful for ADM, but are already reflected in the methods to make systems fairer. These values have in common that they serve the underlying value of fairness. First, ontological heterogeneity is reflected in the efforts that are undertaken to make data sets representative of society as a whole, rather than only focusing on specific (majority) groups. Second, decentralization of power is reflected in the effort that are taken to make ADM more transparent and explainable, demanding accountability and protecting the rights of historically marginalized groups. The discussion of the values of complexity of interaction and applicability to human needs deserves special attention here, because they are inherently related to ADM. Complexity of interaction is reflected in ADM in the methods that are used to design such systems. Machine learning techniques value the complexity of interactions by default, because they are based on complex interactions between features and outcomes. Lastly, applicability to human needs is a core requirement for ADM systems in general. Unlike science, ADM systems are motivated by a direct societal need, making applicability to human needs an essential value for ADM.[10]

The insights about the invention of new values indicate that we may not need new values to guide ADM. Rather, we should focus on making explicit the values that currently play a role in ADM. These values often go unnoticed, because the debate is centered around *bias* rather than *values*. Understanding that a system is biased, for example against women, is a crucial step for recognizing that a system is unfair. However, if the analysis of a system stops here, valuable information gets lost. *Why* do we think it is unfair for a system to be biased against women? Are there situations in which bias would be justified? Is all bias necessarily unfair? Answers to questions like these are often inconclusive, yet give important insights into the values that play a role in a system.

To conclude, in this section I have explained the relation between values, fairness and wrongful discrimination. First, I have argued that value judgments and ADM are inherently intertwined. It is not possible to create ADM systems that are value-free, but this does not mean that all ADM is necessarily unfair. At the same time, I have concluded that value-free people would not guarantee fair AI, because data is inherently social and therefore often reflects societal biases. Secondly, I have argued that values can be used for the better. Indeed, many approaches in the field of fair AI already use this tactic. However, what is currently missing in the debate around fair AI, is the explicit naming of values that play a role in ADM. I believe that this would help our understanding of why and where ADM might go wrong. To help developers and policy makers forward in this pursuit, the next section is dedicated to a taxonomy of values in ADM.

[10] I acknowledge that there could be ADM systems that were not designed to be applied in practice, for example to assess different algorithms in a Computer Science setting. However, I do not focus on these systems in this thesis.

# 5 Taxonomy of Values in Algorithmic Decision Making

To summarize the arguments made in Section 3 and 4 in a more tangible and practical manner, I will propose a taxonomy of values in ADM in this section. More specifically, I will answer the third research question of this thesis:

3. *Where* in the machine learning pipeline do values play a role in algorithmic decision making, *what* values may play a role, and *how* do they play a role?

Note that this section is meant to be descriptive. I do not wish to make normative claims about the values that play a role in ADM. Normative judgments about the topic will be introduced in Section 6 and 7.

## 5.1 Where values play a role

The two main categories with respect to *where* in the ADM pipeline values are used, are (1) in developing a system and (2) in determining the application of a system.

### 5.1.1 Developing systems

In the process of developing ADM systems, there are multiple places where values can come to play a role. The most obvious and important instances of values in ADM designs are discussed in this section. There are, however, other examples of values in ADM designing processes. For instance, in choosing which algorithm to use, values like simplicity or internal/external consistency can play a role.

**Data sets.** The most obvious examples of values that shape ADM are the values that cause data sets to be biased. Values in society shape value-laden data sets. These values are often political or social and have existed a society for a longer period of time. For example, Crawford & Paglen (2021) have showed that a value of physiognomy (i.e., the assumption that people's personal traits are deducible from their looks) have caused image banks to be biased against people who do not conform to beauty standards, including people of color. Values like these often work against historically marginalized groups. What makes matters more dangerous, is that political values shape homogeneous groups of people developing ADM systems, which often causes ADM systems to overlook historically marginalized groups. For instance, (binary) gendered values cause computer games to be more advertised towards men than women, resulting in the fact that more men pursue a career in computer science and AI (Winterson, 2021).

**Feature selection.** Values play a role in determining what features to include in a data set, and what features to leave out. This is an important choice, because ADM systems make decisions about people based on these features. For instance, if one values the principles of luck egalitarianism, individuals' features that are the result of pure luck (e.g., race, gender) should be left out. By contrast, one could also choose to leave out some features that are the result of a person's own choice, for example when someone moves to a certain neighborhood to take care of a family member. Furthermore, choosing particular features for a system implicitly implies a relationship between these features and an outcome. For instance, if one includes someone's race in an ADM system for hiring, one implicitly assumes that race is related to an individual's performance in a job. Thus, we generally should aim to include only relevant features for the particular aim of the system.

Note that simply neglecting sensitive features (e.g., gender and race), often does not lead to the desired result. Machine learning models tend to find a connection between related features that can use as a substitute for the sensitive features. For example, leaving out race as a feature may not lead to fair decisions when a feature of income is included, as race and income tend to be related to each other. It is not my intention to prescribe which features should be left out and which should be included. I believe that it is not even possible to make general claims about this topic, because ADM is context-dependent. Instead, my aim is to make clear that there are many different considerations to be made when it comes to choosing features for ADM. The values we hold, play an important role in this process.

**Fairness evaluation criteria.** Values are used to determine the fairness criteria with which a system is evaluated. The fact that fairness already plays a significant role within AI and ADM reflects a deeper valuation of fairness principles that exist in society. Within this fairness-valuation, choices have to be made with regards to how fairness of ADM practically is ensured. For example, one could aim for high overall accuracy, equal distribution of false positive/negative rates, explainable systems, or representative data sets. These aims reflect different values. For instance, high overall accuracy fits with a valuation of 'faithfulness to facts', while explainable systems fit more with a valuation of transparency and procedural justice. Similarly, equal distribution of false positive/negative rates reflect a value of equality, while representative data sets correspond to a value of equity. Of course, these values and fairness criteria are intertwined and it may not always make sense to treat them separately. My point is that values do play an important role in determining which fairness criteria fit a certain purpose, and in determining which groups need to be protected in certain situations.

### 5.1.2 Determining applications

While the previous subsection discussed values that guide the *development* of ADM systems, we now move to values that are used to determine how a system might be *applied*. The way ADM system are used reveals a lot about the values that certain companies, governments and institutions have. For example, Allhutter et al. (2020) have explained how the application of the Austrian job seeker algorithm reflected several values. The fact that the ADM system makes decisions about people, which in turn directly influences their chances of (re)joining the labor market, is a choice that is influenced by values held by the Austrian government. Clearly, unemployment is seen as a significant problem. Furthermore, the system is designed to maximize efficiency and minimize costs. These characteristics of the application of the system fit with a political values of economic growth, efficiency, productivity and capitalism. The ways in which these values play a role are not necessarily *wrong*, but they do reflect specific choices that are made within a certain political and social landscape.

Likewise, if ADM systems are applied in an ambiguous and unintelligible way (e.g., when law enforcement is reluctant to say how 'risk profiles' for fraud are created), this shows a mindset of secrecy and a lack of trust in citizens. Moreover, it shows that preventing crime is valued, even to the point where the prevention of wrongful discrimination is compromised. The change towards the use of ADM for these kinds of problems also indicates a paradigm shift: we move from ex post to ex ante preventative measures and punishment (McQuillan, 2018). Instead of detecting offenders that law enforcement knows have committed a particular crime in the past, ADM systems that are based on risk profiles can target collective groups of individuals that *might* commit a crime in the future. This comes at the cost of innocent individuals. Thus, the protection of the innocent (majority) weighs less than the prevention of behavior of some. I believe that this says a lot about political values that might exist in a society (e.g., surveillance, crime prevention, willingness to sacrifice privacy and to prevent wrongful discrimination).

These examples show that the way that ADM systems are applied is value-laden. Values determine on what grounds the use of ADM is justified. Moreover, values determine laws and regulations that may prevent or allow the use of a system. For example, while the GDPR allows the use of ADM in businesses (e.g., in banking), it states that individuals should always be able to get an explanation for the decisions ADM makes about them. Furthermore, individuals have the right to consult a human expert as well.[11] Values like fairness and procedural justice guide these rules. However, the fact that systems like the Austrian job seeker algorithm *are* legitimately used, reflect other values that one might not immediately see (i.e., individualism and efficiency).

## 5.2 What values play a role

There are many values which can play a role in ADM. Creating a long list of values will not be the most effective way of demonstrating what values can play a role. Besides, as ADM can be used in an extremely wide range of applications and will likely only develop more in the future, such a list would inevitably be incomplete. Therefore, I will sort the different values based on their general type: (1) computational values, (2) fairness values, (3) social values, and (4) political values. The different types of values interact with each other in many complex ways.

### 5.2.1 Computational values

With computational values I mean values that are used to justify decisions about the computational performance of a system. Performance in this sense is meant to denote how close a system stays to the ground truth (i.e., observed facts in the world). These values come closest to the values that are used to guide scientific research, and include Kuhn's 'standard' values in science: accuracy, simplicity, internal and external consistency, breadth of scope, and fruitfulness (Kuhn, 1977). Furthermore, the standard cognitive values that are mentioned by Douglas (2009) can also be grouped under computational values. Examples include explanatory power and predictive precision. It can be the case that choosing to adhere to one computational value comes at the cost of another. For example, the most accurate ADM systems are often very complex (thereby sacrificing the value of simplicity). Computational values are used to make decisions about the performance of a system, regardless of whether the system is fair.

### 5.2.2 Fairness values

By contrast, fairness values are used to justify the specific fairness criteria that are used in ADM. A helpful way to think about fairness values is to view them as 'schools of thought' about what is fair. For instance, luck egalitarianism is a fairness value or a theory about what is fair. It prescribes that people should be compensated for inequalities that are due to pure luck. This value would guide ADM by excluding individual features that can be assigned to luck (e.g., gender, age, ethnicity). Other fairness values are possible as well. Examples are egalitarianism (focuses on all equal treatment of all people), deontic justice (focuses on accounting for unfairness by making

[11] GDPR Article 4(4) and 22, Recitals (71) and (72).

explicit the historical reasons why certain groups have become disadvantaged), and representational fairness (focuses on equal representation of all groups).

### 5.2.3 Social values

Next, social values are used to guide fairness values. In other words, social values are used to support a specific idea about what is fair/unfair. Of course, fairness is an important social value for ADM, but other examples include equality, equity, dignity, integrity, privacy, transparency, accountability, inclusion and justice. For instance, if one values equality (every individual is given the same resources), they may hold on to an egalitarian fairness value. However, if luck egalitarianism is a more important fairness value, a different social value could be preferred (e.g., equity). Which social values fit best with which idea about fairness is not set in stone, and may depend on the context an ADM system is applied in. The point here is that different valuations of fairness values can lead to different valuations of social values. Thus, while fairness values prescribe what makes a decision fair or unfair, social values prescribe what measures should be implemented in ADM in order to serve the fairness values that are maintained.

### 5.2.4 Political values

Moving from values that are ‘inside’ ADM systems towards values that justify their contextual application, political values are those values that are used to justify ADM within a particular political climate. These include safety, individualism (or collectivism), capitalism (or communism), procedural justice, efficiency, (economic) growth, productivity, crime prevention, democracy, meritocracy, and trust. For example, crime prevention is a political value that may justify the use of ADM for online proctoring systems that closely watch students during exams, aiming to detect cheating. Governments uphold this value by normalizing the surveillance of citizens (e.g., by installing surveillance cameras in streets). Thus, surveillance is a means for crime prevention. Turning towards another example, individualism is a political value that can justify ADM systems, such as the (un)employment algorithm discussed by Allhutter et al. (2020). An individualistic worldview justifies the choice to lay the responsibility by the individual alone, rather than the collective labor market.

Many more examples of political values exist. The main takeaway is that political values are not so much about choices within a system, but more about choices regarding the application of a system. Why do we accept certain ADM systems in society, while we think others are unfair or dangerous? Who is allowed to use ADM (e.g., governments, educational institutions, private companies, individuals, etc.)? What additional rules or regulations do ADM systems need in order to operate fairly? Answering these questions will help recognize the political values that are associated with a particular system. In turn, political values reveal a lot about the potential dangers of a system, the power relation that a system upholds, and the general balancing of values that is used in political processes. For example, online proctoring systems reveal a valuation of crime prevention over privacy, putting emphasis on the wrongdoing of a few rather than protecting the rights of most. Furthermore, we see how governmental decisions (e.g., installment of surveillance cameras) serve as justification for decisions in smaller institutions.

## 5.3 How values play a role

Throughout the discussion of the *what* and the *where*, the different roles of values in ADM have implicitly come forward. In this subsection, I will make the roles more evident, by discussing *how* values play a role in ADM. I will distinguish two main categories, inspired by Douglas (2009): (1) a direct role for values, and (2) an indirect role for values.

### 5.3.1 Direct role

The direct role for values in ADM shapes the design and application of ADM systems in a more explicit manner, rather than serving as tools for guidance. In line with Douglas (2009), a direct role for values in ADM is reserved for values that may in itself be used to justify a decision. Values in this role act as stand-alone reasons to motivate a choice and play a conclusive role. For example, the value of fairness plays a direct role when choosing to impose restrictions on ADM, both when it comes to choices in the developmental stage and choices regarding a system's application. In other words, fairness is often used as a justification *in itself* to design a system in a certain way. This way, a fair model can be preferred over a more accurate model, *because* it is fairer.

Furthermore, the way governments, municipalities, companies and other institutions choose to *use* ADM systems shows clear examples of the direct role for values in ADM. They show that some values are more important for such institutions (or society as a whole) than others. For example, a value of 'preventing crime' is used as reason in itself to justify online proctoring algorithms, that can detect students who cheat during online exams. Furthermore, the values of objectivity, neutrality and efficiency are often used as justification for the appeal and application of ADM as a whole. ADM is valuable for society, *because* it is objective, neutral and efficient. These values are thus directly used as 'arguments' to support the decision to use ADM. If they were not valued as much in society, ADM might not be considered as legitimate as it currently is.

Thus, a direct role for values in ADM directly shapes and guides ADM systems and their usage. Often, such values are implemented by a conscious developer, policy maker, or AI expert. However, this need not be the case. For example, consider a deeply rooted political value, such as efficiency. The use of ADM can be justified *because* ADM systems are efficient, thus letting efficiency play a direct role. Despite their significant and direct role in ADM, deeply rooted values like these may often go unnoticed because they stem from a widely shared 'common sense' that we are all part of. In other words, many political values that shape the way we design and use ADM, guide ADM in a direct but unconscious manner. We usually are so used to these values, that we do not even question them.

### 5.3.2 Indirect role

If values play an indirect role in the development and application of ADM, they are used to weigh the importance, consequences, or motivation for certain choices. Values in the indirect role do not act as justifications in itself, but play a role in the weighing of pros and cons for a choice. For example, in determining which features to include in a data set, an indirect role for values is used to choose which features can legitimately be used as grounds for a decision. When deciding about the inclusion of a feature like race, values such as equality and accuracy are used to reason about the consequences: including a sensitive feature may increase a system's accuracy, but can put equality at risk. A developer has to use values to eventually decide on the topic, essentially suggesting which value is more important.

Contrary to values in the direct role, values that are used in the indirect role do not act as standalone reasons to motivate a choice. Take the above-mentioned example of efficiency. Efficiency often plays a direct role in justifying the application of ADM. However, the fact that we value efficiency can be traced back to a deeper valuation of capitalism in our society. Thus, capitalism plays an indirect role in the choice the apply ADM, by informing our valuation of efficiency. Capitalism is not used as a justification for the decision *in itself*, but it is used as a motivation for specific values we hold, which in turn justifies the decision.

The indirect role for values can be found throughout the entire developing process of ADM, as well as its applications. For example, in determining the (kind of) algorithm that is appropriate for an ADM system, an indirect role for values (e.g., equality, accuracy) is used to weigh the importance of transparency and procedural justice. In determining the rules and conditions under which a certain ADM system can legitimately be applied, an indirect role for values (e.g., procedural justice, efficiency) is used to ensure the system is safe, fair, and effective. They shape the development and application of ADM in the background, rather than guiding it directly.

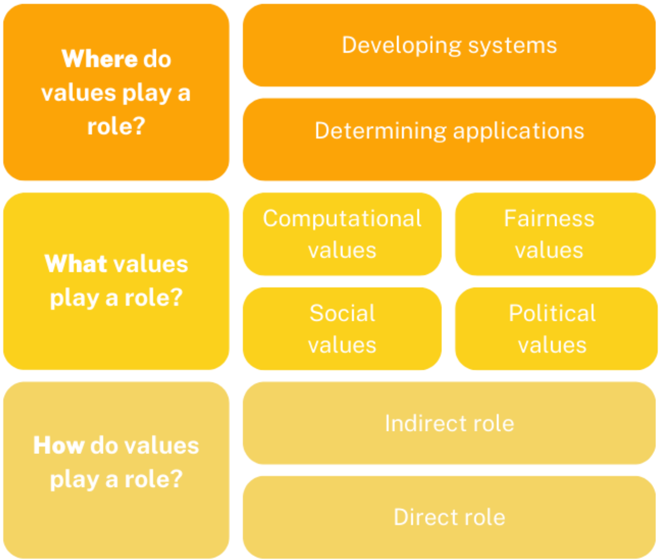

Figure 2: Taxonomy of values in ADM.

In sum, in this section I have proposed a taxonomy of values in ADM. The taxonomy offers a framework to help understand *where* values play a role, *what* values play a role, and *how* they play a role. An overview of the taxonomy is presented in Figure 2.

# 6 Case Study

It should be clear at this point that values *do* play a role in ADM, and it might be a good idea to make them more explicit in the future. This makes it easier to manage the values that play a role in ADM, and helps us see where systems might go wrong.

In this section, I will demonstrate the proposed taxonomy with an existing algorithm that is applied in real-life contexts. More specifically, I will focus on ADM systems that use so-called *risk profiles* for preventative fraud detection. The goal of this section is to provide an example of how the taxonomy can be used in practice. There are two main practical uses of the taxonomy: (1) it can be used by developers and regulators to recognize the values that play a role in ADM systems, ideally resulting in less unintentional outcomes; (2) it can accordingly be used to regulate ADM by informing public sector policies and laws.

I believe that a case-based demonstration of the taxonomy will help the reader understand the use and value of the taxonomy. Furthermore, the context sensitive nature of ADM makes it extremely difficult to develop a top-down approach to gain insights into *all* possible ADM systems. Instead, focusing on a specific case allows us to gain a deeper understanding of the potential issues a system might bring along. This point is emphasized by Algorithm Audit, a nonprofit organization that "builds and shares knowledge about ethical algorithms":

> We believe a case-based and context sensitive approach is indispensable to develop ethical algorithms. One should not expect top-down regulation and legislation to solve all ethical problems in AI and machine learning.[12]

The specific case I will focus on in this section will be Rotterdam's welfare fraud detection system. This system generates a risk score for welfare recipients in the city of Rotterdam. Based on this risk score, recipients can be selected for investigations into their personal lives, aimed at finding fraudulent behavior. This case is relevant for two reasons: (1) fraud detection with risk profiles is a widespread technique for combating fraud, commonly implemented in both the private and public sector (van Schendel, 2019); (2) journalism platforms WIRED and Lighthouse Reports have very recently (March 6, 2023) published a series of extensive stories about Rotterdam's algorithm (Burgess et al., 2023; Braun et al., 2023). Applying the taxonomy to a current case highlights its use in a more effective manner.

The structure of this section is as follows. First, I will provide a general introduction to fraud detection with risk profiles. I will then focus on the specific details of Rotterdam's welfare fraud detection algorithm. Lastly, I will apply the taxonomy to this case, identifying *where* values play a role in the system, *what* values play a role, and *how*.

## 6.1 Fraud detection with risk profiles

The emergence of Big Data has reformed many processes in law enforcement and governmental policies. Big Data has not only made it easier to detect criminal behavior *after* it has happened, but also caused a shift towards a more proactive approach (van Schendel, 2019). Risk profiles play a significant role in this shift. Risk profiling is described as follows (van Schendel, 2019):

> Risk profiling is the categorization or ranking of individuals or groups, sometimes including automated decision making, using correlations and probabilities drawn from

[12] `www.algorithmaudit.eu` (Retrieved March 31, 2023).

> combined and/or aggregated data, to determine the level of risk that is posed to the security of others or to national security by those individuals or groups. (p. 228)

Thus, risk profiles are used to develop a personal 'risk score' for individuals. These risk scores can represent someone's threat to society in a wide range of contexts, for example someone's likelihood to steal, commit identity fraud, or engage in money laundering. The COMPAS system is an example of a system that uses risk profiles to predict whether criminals will re-commit crime (Angwin et al., 2016) . This risk score can be used to inform law enforcement when parole decisions need to be made.

In this section, I will focus on risk profiles that are developed to detect financial fraud. Thus, the risk profiles in this case represent individual risk scores that indicate the likelihood that someone commits financial fraud. Financial fraud is an issue that can have far-reaching consequences for the financial industry and society as a whole. For example, credit card fraud accounts for revenue losses of billions of dollars every year (West & Bhattacharya, 2016). There are many different possible definitions of financial fraud, but for this thesis it will be sufficient to define financial fraud as "the intentional use of illegal methods or practices for the purpose of obtaining financial gain" (West & Bhattacharya, 2016).

Risk profiles are meant to increase the chance of fraudsters getting caught, by checking individuals more effectively and efficiently. Risk profiles indicate which citizens are labeled as 'high risk', based on certain factors that are extracted from the available data (e.g., someone's age, neighborhood or household). These citizens can then be selected for (additional) checks. Risk profiles often are created with self-learning algorithms, that learn a relation between certain characteristics and fraudulent behavior (van Schendel, 2019). This is an example of supervised machine learning, because the algorithm needs access to the features of known fraudsters in order to learn what specific characteristics to look for. *Risk profiling* refers to the act of using risk profiles to make a decision about whether to select a citizen for (additional) checks. This decision can be made by a human (human-in-the-loop), or by the algorithm itself (human-out-of-the-loop).

In essence, risk profiles like these thus distinguish two groups of citizens: one group is selected for (additional) checks, while the other group is not (College voor de Rechten van de Mens, 2021).[13] While we have seen in Section 4 that discrimination between groups is not necessarily unfair or wrong, one can immediately see the potential dangers of risk profiles, for example if profiles are significantly influenced by race or nationality. However, the added value of risk profiles should not be neglected. Financial fraud is a major problem in modern society: West & Bhattacharya (2016) argue that the advancement of modern technologies such as the internet and mobile computing have led to an increase in financial fraud in the past decade. Furthermore, they state that traditional methods for fraud detection are not only time-consuming, inaccurate and expensive, but also quite impractical in the age of Big Data. Additionally, preventing fraud is important because it limits damages that would otherwise need to be paid by innocent citizens. Thus, risk profiles are beneficial for society as a whole, because they make overall processes to detect fraud cheaper and more effective.

In this case study, I will specifically focus on risk profiling systems that are deployed by governmental institutions. Risk profiling in the public sector can for example be used to detect fraud with taxes,

[13] It is also possible that citizens be divided into more than two groups, for example a low, medium and high risk group. The point is that risk profiles discriminate between different groups of people, according to their 'threat' to society.

social welfare and allowances. It is important that fraudulent behavior is tackled in these contexts, because it is crucial that governmental money is well spent, especially in countries with generous social welfare policies. If too many people abuse the system, it will impact the general acceptance of such distributive policies. In the worst case, social policy will need to be reduced and citizens who desperately need benefits will eventually be cut back. This was emphasized in a 2020 court ruling in the Netherlands about the usage of Systeem Risico Indicatie (SyRI), a fraud detection system based on risk profiles. While the system was ruled unlawful because it did not comply with the right to privacy under the European Convention of Human Rights,[14] its purpose was ruled legitimate (van Bekkum & Borgesius, 2021):

> The court noted that "social security is one of the pillars of Dutch society and contributes to a considerable extent to prosperity in the Netherlands." The court added that combating fraud is important, and that it makes sense that the state uses new technologies to combat fraud. (p. 330)

Furthermore, ADM systems come with the promise of being more neutral than human decision-makers. It is thought that ADM has the potential to overcome human biases and limitations. The appeal of risk profiling by governmental institutions can be explained along these lines.

In many bureaucratic contexts, risk profiles serve as decisional aides to human decision-makers (Alon-Barkat & Busuioc, 2023). This is especially the case in contexts where the decisions to be made have a far-reaching impact on individuals' lives, for instance when an individual's access to welfare payments is at risk. Even though risk profiles are usually used to select people for (additional) checks, and do not make a direct decision about the continuation of, e.g., welfare payments, these investigations that recipients with a high risk score are subjected to can be quite invasive (e.g., fraud officers sifting through personal belongings or asking intimate questions about one's love life) (Braun et al., 2023).

Apart from the SyRI case, risk profiling in the public sector has gained a significant amount of (negative) attention in the Netherlands in recent years. In 2018, the Dutch childcare benefits scandal ('Toeslagenaffaire') was brought to public attention. It was revealed that the Dutch Tax and Customs Administration had used a self-learning algorithm to create risk profiles that flagged people that might commit fraud with child care benefits. As a result, tens of thousands of families were flagged and often had to repay the benefits they had already received.[15] This caused major stress and other mental health issues, as well as pushing people into poverty. As it turned out, the system often falsely flagged people as fraudsters. Furthermore, risk profiles were largely based on protected factors, such as nationality.[16] Especially people with double nationalities were targeted by the system.

The Dutch childcare benefits scandal clearly demonstrates the dangers that lie in the usage of risk profiling in the public sector. While risk profiling certainly has advantages and its purpose can be (legally) defended, the potential downsides have far-reaching implications for targeted individuals

[14] The court ruled that the SyRI system did not strike a fair balance between the importance of detecting fraud and the importance of protecting privacy rights. In particular, the court found that the SyRI system and the Dutch government were not transparent enough regarding the system's use of personal data (van Bekkum & Borgesius, 2021).

[15] `https://www.politico.eu/article/dutch-scandal-serves-as-a-warning-for-europe-over-risks-of-using-algorithms/` (Retrieved March 31, 2023).

[16] Factors such as nationality, race, gender, religion and sexual orientations are protected by European law, meaning that they can only in very limited circumstances be used as legitimate grounds for a decision (College voor de Rechten van de Mens, 2021).

and even democratic society as a whole. After all, important democratic values, such as equality, come into questions when people are wrongfully discriminated against.

## 6.2 Rotterdam's welfare fraud detection system

The case that I am focusing on in this section concerns the welfare fraud detection system that was used by the municipality of Rotterdam, the Netherlands. Investigative journalist platforms WIRED and Lighthouse Reports published a story about this system in March 2023, again bringing to light the negative effects that risk profiling in the public sector can have (Burgess et al., 2023; Braun et al., 2023). I chose to discuss this case because of its actuality. Furthermore, while risk profiling is performed by several municipalities in the Netherlands, only Rotterdam has provided information into the algorithm and data used for their purposes. This allows me to go into a more in-depth explanation of the system. While I will deliver critique on Rotterdam's system, I recognize the fact that risk profiling systems of a similar nature are deployed by several municipalities and governments across Europe and North America and I applaud the city of Rotterdam for granting public access to their system.

The city of Rotterdam has roughly 30,000 residents that receive welfare checks (Burgess et al., 2023). Welfare, or *social assistance*, is granted to people that have too little money to get by (e.g., to pay rent, bills, or groceries). Payments are made on a monthly basis, and the amount of assistance someone is entitled to depends on their age and circumstances. To minimize the risk that people abuse the system, it is common for welfare fraud officers to investigate recipients. In 2017, the city installed a machine learning algorithm to make fraud detection more efficient and effective. This algorithm produces a risk score for each of the city's welfare recipients to inform welfare fraud officers when deciding whom to investigate. The system generates a risk score between 0 and 1 for each recipient, 0 meaning low risk for fraud and 1 meaning high risk. This risk score can be viewed as someone's probability to commit welfare fraud.

Each year, the welfare recipients that correspond to the top 10% highest scores are selected for investigation. On average, in the years 2017-2021, this came down to 1,000 people per year (Braun et al., 2023). Reports from these investigations vary from document checks to intensive investigations, including home visits where fraud officers reportedly sift through personal belongings, counting toothbrushes to check how many people live in a home (Braun et al., 2023). Based on the findings of these investigations, the city can hand out benefits penalties, which can consist of fines or the cancellation of welfare benefits completely.

### 6.2.1 Data

The welfare fraud detection system was based on a training set consisting of 315 variables per person (Braun et al., 2023). The training data contained details about 12,707 people who were previously investigated by the city. Half of this group had been found to actually commit fraud (Constantaras et al., 2023). Because the system is based on data that contains ground truth labels, this is an instance of supervised learning. From the data, the system tries to work out relationships between the 315 variables per person, and their likelihood to commit fraud. As stated, the data set was collected by looking at earlier investigations into welfare recipients that the city carried out. Before using an algorithm to select people for investigation, the city used random selection, anonymous tips and category checks that change every year to single out persons. The city of Rotterdam did not disclose information about *why* each person in the data set was selected for investigation (Constantaras et al., 2023).

The 315 different input variables that were used in the system ranged from objective facts, to subjective case worker judgments, to seemingly irrelevant factors. The data set contains a number of variables that are beyond the control of welfare recipients, such as gender and age. Other variables are more controllable, yet are still in a gray area when it comes to determining whether something is a person's 'own choice'. Examples are whether someone is a parent or lives with roommates. Furthermore, some variables are fundamental to why people might need to rely on social welfare on the first place: variables indicating if someone has financial struggles or if someone has a drug addiction (Braun et al., 2023). Variables range from invasive (e.g., the length of the recipient's last relationship), to banal (e.g., how many times a recipient has contacted the city). Other variables seem to be irrelevant for singling out fraudsters, such as whether a recipient plays a sport.

Apart from these (relatively) verifiable facts, the data set also contained more subjective variables (17% of the total amount of variables), which are based on judgments of human case workers (Constantaras et al., 2023). For example, a case worker can note down a comment on someone's physical appearance (e.g., "wears clothes with holes in them"), which results in a binary value of 1 for the variable `comments on physical appearance` (1 meaning "yes, comment", and 0 meaning "no, no comment"). It is not conventional to let subjective judgments like these play a role in ADM systems, because they contradict the claimed neutrality of such systems. After all, data that is based on personal judgments is not value-free, and is thus likely to contain bias (leaving aside for now the fact that even 'objective' measurements may contain bias, as we have seen before). Furthermore, it is important to note that the content of case workers' comments get lost in the data set: all fields were comments *were* made are converted to 1, while all fields were comments were not made are converted to 0 (Constantaras et al., 2023). Thus, if the comment from the previous example were changed to "wears appropriate clothing and looks neat", the value in data set would for `comments on physical appearance` would stay the same.

Some special attention should be granted to the 20 variables that were used to indicate recipients' language skills. In the Netherlands, the amount of a recipient's benefit depends on their proficiency in Dutch.[17] The Dutch Participation Act prescribes that people's right to welfare payments depends on whether they pass a Dutch language requirement. The reasoning behind this being that welfare recipients should provably make an effort to (re)integrate into the labor market and that this is easier if one can speak, write and understand Dutch. The language requirement is met if a recipient has attended a Dutch school or received an integration diploma. If these conditions are not met, recipients can still receive welfare if they improve their Dutch, for instance by taking a course. If recipients are unsuccessful at improving their Dutch despite fulfilling all obligations, their welfare benefits will not be affected. If someone does not meet the language requirement, municipalities are allowed to cut back the benefits or even stop them altogether.

The language requirement is represented as a binary value in the data set (i.e., `language requirement passed/failed`). Apart from this, other variables that indicate recipient's language skills, include information about other their (other) native languages.

[17] `https://www.rijksoverheid.nl/onderwerpen/bijstand/vraag-en-antwoord/wat-is-de-taaleis-in-de-bijstand` (Retrieved April 3, 2023).

### 6.2.2 Model

The model that was used for the welfare fraud detection system, is a so-called *gradient boosting machine* (Braun et al., 2023). The core building block of this model is a decision tree. A decision tree is a supervised learning algorithm, that learns correlations between features in a data set and ground truth labels, and represents those correlations in a tree-like structure. Decision trees are used to guide a decision by specifying a unique path for all individuals. Each path is determined by yes/no questions, indicating if an individual follows the right or left branch next. Eventually, each person ends in a specific 'leaf' (i.e., the end nodes of a tree), which determines their risk score. Figure 3 depicts a simplified example of a decision tree in the welfare fraud detection system.

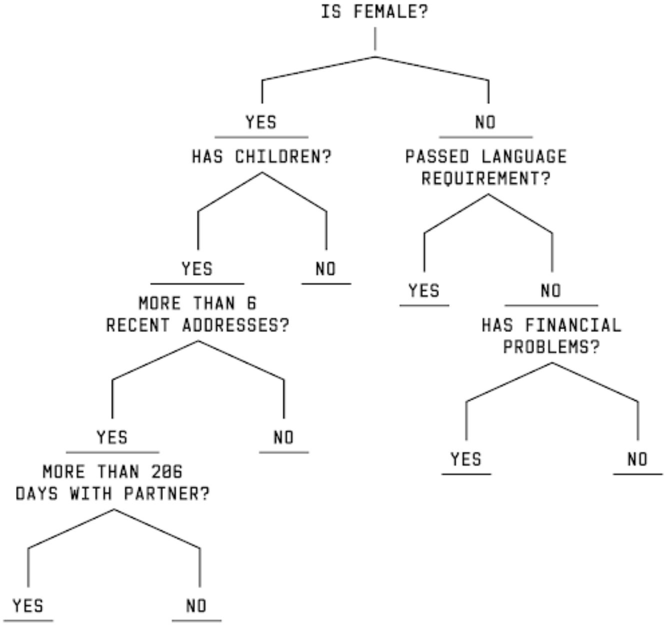
IS FEMALE?
YES
NO
HAS CHILDREN?
PASSED LANGUAGE REQUIREMENT?
YES
NO
YES
NO
MORE THAN 6 RECENT ADDRESSES?
HAS FINANCIAL PROBLEMS?
YES
NO
YES
NO
MORE THAN 286 DAYS WITH PARTNER?
YES
NO

Figure 3: A simplified version of a decision tree in Rotterdam's welfare fraud detection system (figure from Braun et al. (2023)).

The real trees contain more variables and branches. The idea of a gradient boosting machine is to let recipients' variables run through multiple decision trees (500 in Rotterdam's case), each tree learning from the mistakes of the previous tree. This results in 500 risk scores per recipient, which are then averaged and scaled to generate a score between 0 and 1.

### 6.2.3 Evaluation

Internal documents that were obtained by Lighthouse Reports, show that Rotterdam has found the welfare fraud detection system to be 50% more accurate at predicting fraud than randomly selecting people (Braun et al., 2023). Furthermore, documents showed the model's ROC curve. This metric measures the trade-off between people who are correctly labeled as high risk among all people labeled high risk, and people who are wrongly labeled as high risk among all people labeled low risk (Braun et al., 2023). The ROC curve was evaluated by AI-ethicist Dr. Margaret Mitchell, who stated that it indicated that the system was basically random guessing who should be selected for investigation.

Apart from looking at performance metrics, the system was also evaluated based on fairness criteria. Rotterdam's code for the system included a fairness check that tests whether certain groups are over-represented in the 10% highest risk scores the system generates. After all, the recipients in this decile are considered high risk and can get selected for investigation. The fairness definition thus compares the proportion of people who are labeled as high risk, across different groups. According to this definition, a group would be over-represented if more than 10% of individuals belonging to that group would be labeled as high risk. Similarly, a group would be under-represented if less than 10% of individuals belonging to that group would be labeled high risk (Braun et al., 2023). The fact that the city of Rotterdam implemented these checks shows that there were at least some concerns about the undesirable effects the system may have. However, Rotterdam confirmed that the checks were never carried out (Braun et al., 2023).

In 2021, the city of Rotterdam decided to put the welfare fraud system to a halt, after an investigation by the Rotterdam Court of Audit on the use and development of algorithms in the city.[18] The auditor judged that there was insufficient cooperation between the system's developers and the city's fraud officers who used them. Furthermore, the city was criticized because it had not evaluated whether the welfare fraud detection system performed better than the humans it replaced. Lastly, the auditor found that the system had a likelihood of biased outcome, based on the types of features that were used to build risk scored. Since the system had been paused in 2021, the city of Rotterdam has been working on a new and better system, although some have doubts that such a system can be sufficiently transparent and legal (Burgess et al., 2023). Welfare recipients are currently being selected for investigation at random.

## 6.3 Applying the taxonomy

In the next part of this section, I will use the taxonomy proposed in Section 5 to show how Rotterdam's welfare fraud detection algorithm is value-laden. I will follow the same "Where? What? How?" structure as was described earlier. Applying the taxonomy will help us understand the potential dangers of this specific algorithm, as well as understanding how these types of algorithms (i.e., algorithms that use risk profiles to prevent fraud/abuse of the social system) are value-laden in general. As we will see, some problems with Rotterdam's system arise as a result of specific design choices that were made in this case, while others are more structural would continue to exist even if another algorithm design were used. Below, I will talk about the *developer* and *deployer* of Rotterdam's system. With developer I mean the company that designed and built the system (Accenture). The deployer of the system is the city of Rotterdam.

### 6.3.1 Where?

The first step in applying the taxonomy, is determining *where* values play a role in a system. In Section 5, I differentiated between two main categories with regard to where values might play a role: in developing the system and in determining its applications. I will apply the taxonomy to

[18] `https://rekenkamer.rotterdam.nl/onderzoeken/algoritmes/` (Retrieved April 3, 2023).

these two processes in the case of Rotterdam's welfare fraud detection system below.

**Developing systems.** In Rotterdam's algorithm, we can recognize several steps and decisions in the development of the system that are value-laden. They are described below.

- *Data collection.* As previously stated, the training data that was used for the algorithm contained information about 12,707 welfare recipients in Rotterdam. Approximately half of them had been confirmed to commit welfare fraud (Constantaras et al., 2023). The people in the data set were selected for investigations with different selection methods. Rotterdam uses random selection, anonymous tips and category checks (e.g., selecting all men in a certain area) to select people for investigation, but it is not clear how the people in the data set were selected. This might be problematic. Suppose that most men in the data set were selected through random selection, while most women were selected based on anonymous tips from neighbors or case workers. Statistically, this may mean that the women in the data set have a higher chance of actually committing fraud than the men in the data set. After all, anonymous tips could be better indicators for fraudulent behavior than random selections are. The algorithm would in turn learn to associate gender with the likelihood of fraudulent behavior. Values that exist in society cause differences in how people are selected for investigation. For example, the values regarding traditional family roles prescribe that women spend more time doing chores around the house (Perez, 2019), increasing the likelihood that neighbors report anonymous tips about them to the city. This makes women more vulnerable for reporting, because they simply are more visible and can be better monitored than men.

- *Biased data set.* People under the age of 27 were underrepresented in the data set. Specifically, to represent the actual proportion of people in this age range who receive welfare in Rotterdam, there should have been 880 young people in the data set. Instead, there were 52 (Constantaras et al., 2023). This resulted in a bias against young people. This bias could be the result of the fact that a significant amount of the small sample of young people in the data set had committed fraud. However, bias against underrepresented groups can also happen because algorithms will treat these groups as 'outliers' in the bigger group (i.e., not the 'norm'), generally creating less accurate outcomes for these groups. Apparently, the Rotterdam algorithm had learned to associate youth with welfare fraud, based on a sample that was too small to draw any statistically significant conclusions from. Although the unfair outcomes likely were unintentional, the fact that the data set was not representative shows that it was value-laden. Young people were overlooked in the data collection. This probably (and hopefully) was not a conscious choice, but a consequence of the positionality of the developers. For example, overlooking young people might be the result of a team of developers that largely consists of people in a different age range.

- *Feature selection: subjective judgments.* The 315 features that were used to generate risk scores reveal different value judgments. First, subjective case worker judgments make out 17% of features in the system. These judgments mainly address how someone comes across, for example describing their appearance or punctuality. Considering these subjective judgments reveals a certain value-laden standpoint which assumes that one can deduce someone's personal traits from the way someone looks and behaves. For example, one could think that a person is strict and serious because they wear glasses. Moreover, letting subjective judgments like these play a role, is essentially saying that someone's looks and behavior (e.g., punctuality) is correlated with their likelihood of committing fraud, even though the reasons behind someone's appearance or behavior are often unknown. Even if there is a statistical correlation between someone being punctual and their propensity to commit fraud, it would be unfair to automatically suspect people of committing fraud if they are late to an appointment. After all, the reasons behind the lateness are unknown (e.g., someone might be late because they helped an elderly person cross the street).

- *Feature selection: language skills.* Recipients' language skills played a substantial role in determining their risk score. The algorithm contained 20 different features to indicate a person's language skills. It seems to be a logical choice to include features about recipients' language skills, as the Dutch Participation Act prescribes that welfare recipients should have sufficient proficiency in Dutch or make an effort to obtain this. In this sense, committing welfare fraud is indeed correlated with someone's language skills, because one is not (or less) entitled to welfare if they do not speak Dutch. Recipients could thus commit fraud by lying about their proficiency in Dutch. Following this line of reasoning, one would suspect the algorithm to be more likely to flag a person as high risk if they (say they) passed the language requirement and/or has Dutch as native language. However, the opposite is true: recipients who did not pass the language requirement were almost twice as likely to be flagged than people who did not. Furthermore, people who only speak Dutch were over-represented in the low-risk group, while people who speak other languages were over-represented in the high-risk group (Braun et al., 2023).

  Including the language features in the system thus had undesirable effects. The choice to include language features reveals a valuation of the Dutch language and culture, prescribing that welfare recipients should adjust to the Dutch culture. Normative statement like these are not neutral or objective. Furthermore, while checking the language requirement in the system might be a straightforward choice based on the Participation Act, one can question the use of the other language features. These features seem to suggest that one is more or less likely to commit fraud based on which languages one speaks (besides Dutch). Including these features is not obvious and reflects the positionality of the developers of the system, which may have led them to believe that (foreign) language skills and fraud propensity are related.

  Again, even if this relation was demonstrated, there are legitimite reasons *not* to include language features. College voor de Rechten van de Mens (2021) objected against the use of language skills in the system, because they are proxies for ethnicity and nationality. After all, the chances of someone failing the language requirements is much bigger for people who were born outside of the Netherlands than they are for someone who was born in the Netherlands. Furthermore, having another native language than Dutch is a proxy for nationality and ethnicity as well.

- *Feature selection: amount of variables.* Rotterdam's system contained 315 variables, some of which seem irrelevant to the task. A commonly heard argument for including as many factors as possible, even if these factors seem to be potentially discriminatory (e.g., language skills), subjective (e.g., case worker judgments) or irrelevant (e.g., whether someone plays a sport), is that more data will only improve the system and its objectiveness. After all, if one *does* make a decision about which features to include and which to ignore, one uses value judgments to do so. Staying clear from these decisions thus means staying neutral with regard to feature selection. However, I argue that this is still a value-laden decision, especially in light of the potential harms it may involve. As is already stated, every decision that leads to a feature being used in a system, is essentially a decision about the relationship between that feature and a risk score (or whatever else a system may produce). These are not obvious facts, but instances in which values are weighed against each other (e.g., "Do I include gender, which might improve accuracy at the cost of fairness?"). Furthermore, it may be tempting to think that including as many features as possible will increase a system's performance. After all, having access to more data means that a system can make more informed decisions. However, adding more features also increases the risk of finding correlations that are meaningless, unstable or undesirable. It becomes almost impossible to check which correlations are legitimate and which are not, thereby decreasing the reliability of the system.

- *Decision tree.* We have already established that the choice for using decision trees reflects a valuation of transparency and simplicity. Here, I would like to zoom in on the decision tree that is displayed in Figure 3. Taking a closer look at the tree, it becomes apparent that it exposes certain values in itself. Dependent on someone's gender, a person gets asked different kinds of questions that ultimately lead to their risk score (leaving aside for now the fact that the tree only recognizes a binary gender concept). If one is female, questions are centered around one's domestic situation (e.g., if one has children or a partner). However, the decision tree evaluates men based on criteria concerning their language skills and financial situation. My point is not to show that this divide is fairer/unfairer for men/women, which it may very well turn out to be. Instead, my point is to show that the divide in questions is not neutral, but rather shaped by societal values about how we judge women versus men. These values are represented in the data set, which explains why we see them in the learned decision tree as well.

- *(Fairness) evaluation criteria.* The system was evaluated based on its accuracy and the ROC curve, which plots the true positive rate against the false positive rate. These metrics suggest that the developers/deployers (it is not clear who carried out the evaluation) prioritized the system's faithfulness to facts. This reasoning can be explained by the common belief that objectiveness implies fairness. Ironically, the ROC curve showed that the system did not perform much better than random selection. It is unclear if this was known to the developer/deployer, or if they knowingly decided to use the system anyways, possibly justifying the use by pointing to its (cost) efficiency.

  To ensure the system operated fairly, a fairness check was installed (although not carried out). This check reveals if certain groups are over-represented in the high risk group. As discussed in Section 6.5.1, this check reflects a valuation of equality: every individual, regardless of what groups one belongs to, should have the same 'starting position'. It is not fair if people are already disadvantaged from the start, just because they belong to a certain group. It is up for debate how valuable this fairness evaluation is. After all, it may be the case that the baselines for different groups differ, such that, for example, women *do* commit welfare fraud more often than men (ignoring for now the possibility that this statistic is the result of selection or confirmation bias, or specific circumstances in society). It is beyond the scope of this thesis to discuss the use of Rotterdam's fairness evaluation at length, but it should be clear that choosing for a specific metric (and thereby leaving others behind) reflects certain value judgments in itself.

**Determining applications.** Apart from values that were used in the development of Rotterdam's system, values also played a role in determining how the system was applied. These considerations are described below.

- *From ex post to ex ante measures.* The application of Rotterdam's algorithm can be situated into a bigger debate about the paradigm shift of Big Data, both in the public and private sector. McQuillan (2018) argues that this shift is characterized by a move from ex post to ex ante preventative measures and punishments. Thus, instead of focusing on only zooming in on the people that have already committed fraud, ADM and Big Data treat *everyone* as a potential wrongdoer: all welfare recipients receive their own risk score that indicates how trustworthy they are. This view goes hand in hand with an attitude of distrust and suspiciousness, moving from a mindset of 'innocent until proven guilty' to 'guilty until proven innocent'. Not surprisingly, WIRED and Lighthouse Reports refer to Rotterdam's welfare fraud algorithm as the 'suspicion machine' (Constantaras et al., 2023).
- *Algocracy.* The increasing use of algorithms in the public sector leads to an algocracy: a system where algorithms collect, sort and organize the data that is used to make decisions with (Danaher, 2016). This change is commonly justified with arguments referring to time/cost efficiency, accuracy, and objectiveness or neutrality. These values are thus deemed important in our society. Danaher (2016) highlights the 'threat of algocracy': the opacity of algorithmic governance systems pose a threat to the legitimacy of public decision making processes. The use of algorithms in the public sector is especially dangerous because the opportunities for human participation and understanding are limited. This threat can be recognized in Rotterdam's system, where it was nearly impossible for recipients to object against a decision.
- *'Scientific' relation between language skills and propensity to fraud.* As mentioned above, including language features can be explained by the Dutch Participation Act, but the system's bias against recipients that did not pass the language requirement contradicts this explanation. Closely related to the values that play a role in the feature selection of the system, Rotterdam's algorithm allows welfare officials to state that people who do not pass the language requirement are untrustworthy or have a higher propensity to commit fraud. This point is stressed by Constantaras et al. (2023), who state that the system gives the illusion that there is a scientific association between poor Dutch skills and fraudulent behavior. I believe that this illusion stems from the promise that the algorithm is neutral and objective. Because objectiveness, neutrality and science are highly regarded in our society, systems such as Rotterdam's welfare fraud detection algorithm are justified.

### 6.3.2 What?

The second step of the taxonomy is to identify *what* values play a role in the system. Recall that I distinguished four types of values: computational values, fairness values, social values and political values. I will single out each of these types of values below.

**Computational values** are used to make decision about the performance of a system, regardless of whether the system is fair. Computational values tell us why the inner workings of a system are justified. The following computational values can be identified in Rotterdam's welfare fraud detection system:

- *Accuracy.* Internal documents showed that the city of Rotterdam reported that the system was 50% more accurate at predicting fraudulent behavior than randomly selecting people for investigation (Braun et al., 2023). Although the disclosed ROC curve contradicts this, the fact that Rotterdam used the system's accuracy as a justification of the use of the system, shows a valuation of accuracy. It is not clear if Rotterdam used accuracy scores to inform decisions in the designing phase, or they were used to reason about the system ex post. However, from the repeated reporting of accuracy scores it can be concluded that accuracy played a role in the justification of the system.

- *Interpretability and simplicity.* The AI model that was used for the system was a gradient boosting machine, whose building blocks are decision trees. This model is often regarded as relatively simple and interpretable, especially compared to 'black box' models such as deep neural networks. Even though Rotterdam's model takes into account 315 data points and generates 500 trees for each individual, this is relatively intuitive compared to the billions of nodes some deep neural networks are made up of. Furthermore, decision trees offer an intuitive explanation of why the system makes certain decisions: for each tree, one can simply follow the path that leads to the ultimate end node, corresponding to the decision. Considering the fact that most models that are more complex and harder to interpret often achieve higher accuracy scores, it is likely that simplicity and interpretability were important for the system's developers and its deployer.

- *Speed.* Although there is no openly available reporting of the system's time efficiency, it is likely that it played a role in its justification. ADM systems are regularly praised because they can make decisions much faster than humans can. This in turn makes processes much cheaper. Thus, speediness in many cases also reflects a valuation of cost efficiency.

**Fairness values** are used to justify the fairness criteria that are used in a system. As such, fairness values represent 'schools of thought' on what is fair. The following fairness values can be identified in Rotterdam's welfare fraud detection system:

- *Negligence with respect to fairness.* First of all, the fact that a fairness check was implemented but not carried out, shows a lack of focus on fairness. Of course, we do not know *why* Rotterdam did not carry out the fairness checks, and there can be a reasonable thought behind it. Nevertheless, the fact remains that Rotterdam did not perform any checks with regard to the system's fairness, indicating that fairness might not have been an important (enough) value. This example shows that applying the taxonomy to a case can also reveal something about values that are missing in an application.

- *Egalitarianism.* If we take a look at the fairness check that was implemented (but not used), we see a value of egalitarianism. The check measures whether certain groups are over-represented in the highest decile of risk scores. The underlying thought of this metric is that it is unfair if people have a much higher chance of being labeled high risk, 'just because' they belong to a certain group.[19] This view corresponds to egalitarianism, because it indicates the importance

[19] The check that was installed did not account for correlated variables, so it is not entirely possible to check if someone receives a high risk label 'just because' they belong to a certain group. For example, the system might give higher risk scores to women compared to men. It could be the case that women are disadvantaged 'just because' they belong to this group, but it could also be that being a women happened to correlate with a younger age or other variables that increase one's risk score.

of equal chances for every individual. For instance, a black person should, all other factors being equal, have the same chance of being labeled high risk as a white person would. Contrary to luck egalitarianism, variables that are due to pure luck are considered in the model (e.g., gender, age, native language).

- *Procedural justice (i.e., right to a fair procedure).* Several choices that were made with regard to the model itself and its application, suggest that the developers and deployers of the algorithm value recipient's right to a fair procedure. This is mainly reflected in the choice for a relatively simple and interpretable model (decision tree) and a human-in-the-loop design. These measures can be seen as a safeguard for a fair procedure for the individuals that are targeted by the system, by offering an explanation of made decisions in human terms. Institutions can use these safeguards to check whether the system operates fairly and to object against decisions that were made by the system.

**Social values** are used to support a specific idea about what is (un)fair. Social values can guide fairness values, by determining what values are important for a specific application, target group or institution. The following social values can be identified in Rotterdam's welfare fraud detection system:

- *Equality.* In line with egalitarianism as 'school of thought' about what is fair, equality seems to be an important value for Rotterdam's system. Equality is a social value (and not a fairness value), because equality is used to guide decisions in the designing process. If one values equality, egalitarianism is a suitable 'school of thought' on what is fair. An egalitarian view in turn determines whether certain aspects of the system are fair and in which circumstances they are fair. The fairness check that was installed, tests whether different groups are treated equally with respect to the probability of being labeled high risk. Equality is arguably one of the most important values in a democratic society, so it may not be a surprise to find this value in a system that is deployed in the public sector.

- *Transparency.* Decision trees offer some amount of transparency, especially when compared to complex models such as deep neural networks. Decision trees provide intuitive visualizations of the decision making process. Although the calculations that were performed in order to *make* the trees are more opaque, decision trees at least can explain the steps that lead to a specific decision. It is important to note that while the city of Rotterdam had access to the decision trees that were generated for the system, the people who were targeted by them did not. As a result, objecting against a decision was nearly impossible for people who were labeled as high risk. Thus, while the chosen model was relatively transparent to the people who had access to it, it was opaque to people who were targeted by it.

- *Accountability.* Rotterdam emphasized that the welfare fraud detection system did not make decisions about recipients' welfare payments on its own. The city chose for a human-in-the-loop design, always making sure that a human assessor looks at the risk profiles that were generated. Furthermore, the system 'merely' indicated who should be selected for extra investigations (performed by humans), making no statement about the amount of welfare money someone should receive. By installing these safeguards, Rotterdam took (at least some) accountability for the system.

**Political values** are those values that are used to justify a system within a particular political climate. On what grounds is a system justified, and what makes its decisions legitimate? Political values tend to focus on how a system is applied, but can also impact the inner workings of a system. The following political values can be identified in Rotterdam's welfare fraud detection system:

- *Crime prevention.* As the sole purpose of the system is preventing welfare fraud, which is criminal behavior, it is quite obvious that the system is justified because crime prevention is valued by the city of Rotterdam (and bigger governmental institutions). If preventing crime were less important, the system might not have been developed and deployed in the first place.
- *(Cost) efficiency.* The strong appeal of risk profiles is that they make fraud detection more efficient and effective. Even though it has become clear in recent years that risk profiles come with certain dangers, especially their tendency to harmfully discriminate against a certain group, the use of risk profiles is justified by their unprecedented efficiency. AI in general is able to perform much more complicated calculations than humans ever could, taking into account thousands of factors that might play a role in determining a risk score. It follows that systems that use risk profiles simply are able to (potentially) catch significantly more wrongdoers than human assessors would. Whether this actually happens or not, this argument is often used as a legitimate reason to use risk profiles, leaving the potential harms behind. Furthermore, using a system like Rotterdam's is significantly more cost-efficient than letting human assessors determine which individuals to select for extra checks. After all, humans have to be paid and only work a limited amount of hours, while an ADM system is much cheaper (once it has been developed) and can work around the clock. The valuation of cost efficiency is expected in a capitalist society, where it is common to measure a product's purpose in terms of economic worth (i.e., how much money the product saves/brings in).
- *Objectivity.* Like most ADM systems, the Rotterdam's welfare fraud detection system is deemed legitimate because it claims to be more objective than human assessments. This reflects a valuation of objectiveness and neutrality. As has been discussed throughout this thesis, a claim of objectivity is often used as justification because "objective statements cannot be unfair". As opposed to subjective judgments, which are inherently shaped by human bias, objective claims are thought to be neutral and superior. However, as I have explained, objectivity is no guarantee for fairness and neutrality. In light of the use of objectivity as a justification of the system as a whole, it is surprising to see the use of subjective factors inside the system. Subjective judgments about recipients, made by case workers who have interacted with them, play a role in the calculation of their risk scores. This directly contradicts the claimed objectivity of the system. Here, we thus see how applying the taxonomy can expose contradictions and inconsistencies within a system. This way, deployers of the system (e.g., the city of Rotterdam), can assess their own reasoning behind a system and recognize topics or parts that need to be improved or accounted for.

### 6.3.3 How?

The last step in the taxonomy is determining *how* values play a role in a system. Following Douglas (2009), I identified two roles that values can play: a direct and indirect role. The different roles that values play in Rotterdam's welfare fraud detection algorithm are described below. It should be noted that it is difficult to indicate the ways in which values were used to guide decisions in Rotterdam's case, because I do not have access to the specific thought processes behind them, nor to the documents with which Rotterdam commissioned the system or how the developers were instructed. I have therefore tried to deduce the most obvious instances where a(n) direct/indirect role for values was used.

**Direct role.** A direct role for values in ADM is reserved for values that may on itself be used to justify a decision about a system (which can be either a design choice or a choice about how to apply a system). This often happens consciously, but sometimes the values that are used in a direct manner are so ubiquitous that decision makers use them to justify their decisions without them noticing or questioning it.

- *Justifying the use of ADM.* The decision to turn to ADM to combat welfare fraud is justified by one value in and of itself: objectivity. Following this reasoning, ADM is a worthwhile endeavor because it offers objective and fact-based support for decision making. The value of objectivity is used as a reason on itself for this argument, and thus is used in a direct manner. This fact may easily be overlooked, because our valuation of objectivity goes without saying. However, objectivity often plays a legitimate direct role in the justification of a method or system. Moreover, efficiency also plays a direct role in justifying ADM systems. Because such systems often are way more efficient, in terms of both cost and time, their use is sufficiently motivated.
- *Justifying ADM to prevent crime.* Crime prevention is in and of itself used to justify Rotterdam's system. In other words, it is because we, as a society, value crime prevention, that we approve of systems such as Rotterdam's welfare fraud detection algorithm. The value of crime prevention is used in a direct manner, because it is used on itself to justify the decision to use ADM systems for this purpose. Again, the fact that crime prevention is a certain *value* and is not some objective, obvious goal, might be overlooked. However, one could imagine a scenario or society where crime prevention is less important and the use of Rotterdam's system would not be accepted, for instance because the threats it poses on legitimate democratic processes are deemed too dangerous. This example shows that the seemingly 'obvious' values we hold are not indisputable.
- *Justifying the use of decision trees.* In preferring decision trees over other methods to build AI, the value of transparency and simplicity may play a direct role. It is sufficient to choose decision trees, solely because they are more transparent or simpler. Thus, these values are used in a direct manner, and provide legitimate grounds for a decision.

**Indirect role.** If values play an indirect role in the development and application of ADM, they are used to weigh the pros and cons of a choice, or to weigh the importance or consequences of a model. In the following aspects of Rotterdam's system, one can recognize an indirect role for values:

- *Performance evaluation.* The choice to apply certain evaluation metrics of the system requires an indirect role for values to weigh the consequences of the errors that the system might make. In Rotterdam's case, developers and deployers chose to measure the system's performance by evaluating its ROC curve, which plots the true positive rate against the false positive rate. An indirect role for values is reserved for this choice, to reason about the kind of mistakes the system might make and what would be the consequences. In this case, a false positive would indicate a person that is labeled as high risk, but does not commit welfare fraud. Given the fact that the investigations into people with a high risk profile can be extremely intrusive, false positives have negative consequences for the concerned individuals. On the other hand, the true positive rate indicate the percentage of people that were labeled high risk by the system, who actually commit fraud. A high true positive rate is thus essential for an effective system. Looking at this context, Rotterdam's decision to look at the ROC curve becomes more clear: one would like to minimize the false positive rate, while maximizing the true positive rate. While this is almost always the case, other measures could have been used to evaluate the system too. For example, one could look at false/true negatives as well. Arguably, the prevention of false negatives (i.e., people that are labeled as low risk but do commit fraud) is important as well. Every system suffers from a trade-off between false positives and false negatives. Rotterdam chose to focus on false positives only, which is a choice guided by value judgments.

- *Fairness evaluation.* The code included a fairness check, which tests whether certain groups are over-represented in the high risk group. An indirect role for values is used to make decisions regarding which fairness checks to include, for example to determine what fairness definition is appropriate in a certain context. In Rotterdam's case, we see that the developers/deployers found it important that people from different groups have equal probabilities of receiving a high risk score. As was discussed above, this reflects an egalitarian fairness value. Other fairness checks could have been installed as well. For example, one could compare the ROC curve of different groups. This could expose certain biases, for example when the false positive rate for one group is significantly higher compared to another group. My point here is not to criticize Rotterdam's decision with regard to the fairness evaluation metrics they chose to install, but rather to point out that there is an indirect role for values that help developers and deployers to choose what is fair in a certain context, whether negative outcomes for individuals outweigh negative outcomes for bigger groups, and which people should be protected.

- *Feature selection.* By choosing which features are used in the system, an indirect role for values is reserved for weighing the influence of a feature on the outcome (i.e., someone's risk score) against the potential downsides of using this feature. For example, take the features that are used to denote a recipient's language skills in Rotterdam's case. It is reasonable to expect that features that have to deal with someone's language skills are correlated with so-called 'protected features', such as nationality and ethnicity. Nevertheless, Rotterdam chose to include the language features, possibly because they also show some correlation with propensity to fraud. This choice was made with an indirect role for values: the positive and negative consequences of including language features were weighed against each other, and it was decided that the positive consequences outweigh the negative. The same weighing could have occurred with every feature that was included or eventually left behind. These decisions are not 'objective' or 'neutral', as it is not obvious which features to include. Decisions like these need to be guided by values of the developer, deployer, or society as a whole.

# 7 Preliminary Conclusions

In this section, I will draw conclusions from the case study in Section 6. The conclusions I draw lead to some more general and practical advice as to how to apply the taxonomy. Furthermore, I will point to the limitations of the taxonomy.

## 7.1 Conclusions from case study

The case study, focused on Rotterdam's welfare fraud detection algorithm, has served as an example of how the taxonomy of values in ADM can be applied to real-life cases. Applying the taxonomy to the case has uncovered the values that play a role in the system. As a result, it is easier to understand on what grounds the application of system was justified by the city of Rotterdam, and what were the dangers of the system. There are four main findings of the case study. The first and second main findings refer to specific choices that have been made in the designing process of the system. The third and fourth main findings concern the application of the system; these findings would not cease to exist if Rotterdam would design a different system for the same purpose.

First, the taxonomy helped us see that the model makes value-laden judgments based on nationality. The use of language features in the system may be explained by the Dutch Participation Act, which prescribes that welfare payments depend on a recipient's fluency in Dutch, or effort to learn Dutch. The Dutch Participation Act reveals a judgment that values (re)integration into the labor market. Furthermore, including these features caused the system to discriminate based on language skills. People who speak certain (second) languages, like Turkish, receive higher risk scores than others. One can easily see how language skills are correlated with 'protected' features, such as nationality and ethnicity. Even if it were statistically proven that language skills correlate with fraud propensity, it is a value-laden choice to include these features in a system like Rotterdam's. After all, machine learning algorithms cannot distinguish correlation from causation: they make predictions that are based on correlations found in the data, but we treat these predictions as causal inferences. Including language features is thus another way of saying that it is legitimate to treat *all* people with certain language skills differently than others, by assigning them higher risk scores by default.

Second, the taxonomy exposed value-laden judgments about gender. The decision trees that were generated reveal a difference in the risk-calculation between women and men. Where women are evaluated based on traditional 'domestic' issues, like relationships and children, men are judged based on their job, language skills and financial situation. This is a direct reflection of the gendered division in society: women generally perform more unpaid caring and domestic tasks compared to men. Decision trees are shaped by algorithms, so the fact that they reflect gendered value judgments is most likely not caused by AI developers directly, but rather by hidden values in the data set. Indeed, we have seen that the data collection was not neutral with respect to gender. As it is unclear what method of reporting was used for the recipients in the data set, it cannot be ruled out that women are negatively affected by the way recipients were selected. Because women are more visible to neighbors and typically have more interactions with people in their vicinity (e.g., school teachers, store employees, parents, doctors, etc.), they are more susceptible to anonymous tips being reported about them. Given that anonymous tips generally have a higher true positive rate than, say, random selection (Braun et al., 2023), there is a substantial chance that women were disproportionately disadvantaged by the way data was collected.

Third, the taxonomy revealed that the general justification of the Rotterdam welfare fraud detection system is largely based on a valuation of crime prevention. In other words, we find preventing crime important. This seems logical: criminal acts typically harm innocent people, and this harm should be prevented wherever possible. In Rotterdam's case, the motivation behind crime prevention is primarily financial: welfare fraud supposedly costs society (and citizens) money that could be used for other valuable purposes. Furthermore, welfare fraud is detrimental for a generous welfare policy. Thus, the taxonomy reveals a valuation of economic prosperity and growth as well. This fits capitalist countries such as the Netherlands (although, of course, countries with other economic systems also benefit from limiting unnecessary spending). Furthermore, the taxonomy exposed that Big Data and AI treat everyone as a possible suspect: every welfare recipient gets assigned a risk score, on the basis of which officials decide whether or not to investigate them. In such investigations, recipients have to prove that they did/do not commit fraud. Accordingly, this might cause a shift in criminal law, moving from 'innocent until proven guilty' to 'guilty until proven innocent'. Concluding, we see how Rotterdam's system is a direct product of a valuation of crime prevention, and how this value relates to others (e.g., economic prosperity and distrust). These insights help us see the way that values create a ground for justification of the system, as well as recognize potential dangers that go along with them (e.g., disruption of the justice system). For example, if the drive behind the system had less to do with financial gain and more with justice, the system's focus might shift from detecting possible fraudsters *in general* to detecting the possible suspects for a *specific* known case of fraud.

Fourth, the taxonomy brings to light the valuation and immense importance of objectivity and efficiency in society. The use of Rotterdam's welfare fraud detection system is justified by pointing to its efficiency in terms of cost and time, and its predictions are deemed legitimate because of the data-driven approach and the objectivity that supposedly follows from this. Algorithms are cheaper and faster than humans, and can reason with much more parameters. Furthermore, algorithms reason on the basis of thousands of real-world data points. These data points surely seem to be more objective than the individual reasoning a human officer might use to predict recipients' fraud propensity. After all, human reasoning is influenced by value judgments. In light of these arguments, it seems common sense to implement systems like Rotterdam's welfare fraud detection algorithm. However, it is not that obvious: as a society, we *choose* to value objectivity and efficiency. It could be different. In a society where efficiency is less important, one might not want to risk implementing discriminatory systems, even if they are more efficient than human methods.

Furthermore, the case study revealed some general uses for the taxonomy. First, the taxonomy can be used to expose inconsistencies and contradictions in the value judgments that are contained in a system. For example, the use of subjective judgments as features in the system refutes the general valuation of objectivity that we have found throughout the system. Second, the taxonomy can be used to expose values that might be expected in a system, but are missing. For example, even though fairness checks were implemented, Rotterdam did not carry out any fairness checks to evaluate the system. Although we do not know the reason behind this choice, it *is* a choice to not report anything about the system's fairness scores. Especially in a time where the concerns about fair AI and the dangers of risk scores are so commonly articulated that it is hard to believe that Rotterdam simply 'forgot' about the fairness evaluation of the system.

## 7.2 Practical recommendations

The conclusions from the case study, together with my personal experiences of applying the taxonomy to the Rotterdam case, lead me to give some general recommendations and practical advice as to how to apply the taxonomy. I will dispense this advice here.

1. Pay special attention to values that are easily overlooked because of their ubiquity. These values, such as efficiency, objectivity, and economic growth, might be self-evident, but will often reveal a lot about how, where and why a system might be deployed and applied. Making these values explicit reveals the trade-off between values. For example, ADM systems that use risk profiles often involve a trade-off between crime prevention and fairness, in the sense that the justification of a system relies on weighing the importance of catching wrongdoers against the importance of preventing the implementation that could possibly discriminate against (groups of) people. Analyzing the trade-offs between values does not only help one obtain certain insights into society, but also helps one recognize potential dangers and pitfalls of a system.

2. Keep in mind that the main purpose of the taxonomy is merely to expose the value judgments that might be embedded in ADM systems. The taxonomy is thus *not* designed as a tool for generating constructive critique. Moreover, some values in ADM systems can be legitimate, depending on the context and application. My main point is that, in the process of applying the taxonomy itself, it is not necessary to form opinions about a system. Of course, *after* the taxonomy is applied, one has to reason about the desirability and legitimacy of the values that are embedded in a system, and make adjustments where needed (either in the system or in how it is applied). For example, in our case we see that some values should be managed, to prevent the undesirable discrimination based on nationality and gender. The *How?* section of the taxonomy was designed to support reasoning about the legitimacy of value judgments in a system. Determining how values play a role (direct or indirect), may help in deciding whether these values are legitimate.

3. Do not take the taxonomy too literal. The taxonomy does not have to be applied word-for-word, the way I have described it in this thesis. For example, I have found it to be helpful to start with the *Where?* section, but this may be different for other people or with other applications. Going through the different steps of the taxonomy will likely reveal general patterns of value judgments throughout a system. It is most helpful to view the taxonomy as a tool for guiding questions about a system. It may not always be necessary to apply the taxonomy in the same order or intensity as I did. Preferably, let multiple people with different backgrounds apply the taxonomy on the same case. Furthermore, the taxonomy can be used by people in different roles, such as designers, developers, employers, and third parties assessing algorithms. A multitude of perspectives will decrease the chance that an important point is overlooked.

## 7.3 Limitations

While applying the taxonomy, I also encountered a few limitations. First, I am well aware that, more often than not, one does not have access to the internal workings of a system. It is therefore often impossible to see what value judgments were used for decisions with regard to the internal design or feature selection of a system. Practically, this may mean that it is not possible to apply the taxonomy in its entirety. While this fact indeed may limit the usefulness of the taxonomy, certain parts of the taxonomy remain applicable to systems of which the internal choices are unknown. As was revealed in the case study, the taxonomy can effectively be used to expose the value judgments that are used to justify the use of a system. These insights are valuable, as they reveal underlying structures in society and help determine where the dangers of a system lie.

Second, a limitation of the taxonomy is that it is by no means complete. In the *Where?* section, many more places where values can play a role could be mentioned. In the *What?* section, many more examples of values could be named. In the *How?* section, many more roles of values could be described. However, the taxonomy remains helpful as a tool to guide questions about a system. Following the structure of the taxonomy and asking the *Where? What? How?* questions, will reveal the value judgments that underpin a system. Furthermore, as the values that shape a system depend significantly on the type of system and its use, I am convinced that a complete taxonomy is not possible to create, and also not useful. Whoever applies the taxonomy might come up with different answers and even new categories. Thus, incompleteness is as much a limitation as it is a strength. For the taxonomy to be useful, it needs to be flexible and open to different interpretations.

Third, the taxonomy focuses on exposing value-ladenness of ADM systems, rather than the interventions that might be needed to make systems better or fairer. It is thereby not enough to only apply the taxonomy; the lessons learned from this practice will need to be converted to actions. However, it was my purpose here to show that ADM is not neutral and contains value judgments. The case study has showed that the taxonomy is capable to do this.

# 8 Discussion

It is now time to take a step back and see how the taxonomy relates to the bigger debate about values in AI, fair AI, bias and scientific objectivity. Let us recall the three research questions that guided this thesis:

1. How does the value-neutrality of science relate to the claimed objectivity of algorithmic decision making?
2. To what extent can the philosophical arguments *against* value-neutrality of science be applied in the context of the objectivity of algorithmic decision making?
3. *Where* in the machine learning pipeline do values play a role in algorithmic decision making, *what* values may play a role, and *how* do they play a role?

In this section, I will provide an answer to these questions. Furthermore, I will explain the relation between the taxonomy and some of the other approaches that we see in the field of fair AI.

## 8.1 Answering the research questions

The first half of this thesis focused mainly on the first two research questions: 1) How does the value-neutrality of science relate to the claimed objectivity of ADM?, and 2) To what extent can the philosophical arguments against value-neutrality of science be applied in the context of ADM? In Section 3, I tried to answer these questions. More specifically, I explored the relation between the value-neutrality of science and the value-neutrality of ADM, by applying arguments from the philosophy of science to the field of ADM. I concluded on two main similarities and two main differences between these fields.

The first parallel between science and ADM was found in the epistemology. Both science and ADM can be characterized by a 'hunger for objectivity'. It is presumed that there is a greater, objective order to the world that we cannot access directly. Instead, a data-driven approach is needed to obtain these objective truths. This approach needs to be as objective as possible, meaning that there is little to no place for value judgments. Both science and ADM make inductive leaps from observed to unobserved data: we learn something from observations in the past, so we can predict the future. However, there is a major difference between the fields when it comes to the inductive leaps made. In science, the learning is mediated by theories, so that observations are used to develop, calibrate, test, and falsify theories. Theories in turn are used to make predictions. Conversely, AI and machine learning directly use past data to predict future data, without the mediation of a theory that is used to contextualize observations.

Furthermore, a parallel that relates to epistemology refers to the dangerous hierarchy between the 'self' and the 'other'. In the philosophy of science, this idea has been discussed by Longino (1995) and Halpin (1989), who state that science is not value-free because the 'self' (i.e., the collective body of researchers) does not represent *one* objective standpoint. Instead, the 'self' consists of multiple people, all shaped by different values and social contexts. Claiming that this plurality does not exists, allows for oppression of the 'other', because the scientist distances oneself from the object of study (which automatically becomes the 'other'). These arguments transfer to ADM (and, in particular, to feminist AI), where Adam (1995) argues that the endeavor to create machine intelligence is based on a prioritization of a narrow concept of knowledge, typically assigned to the 'self' of science: white, heterosexual, Christian, middle-aged males, belonging to the middle or upper class. Furthermore, Lepri et al. (2018) state that the widespread use of ADM could lead to a situation where a powerful majority have access to tools and knowledge of systems that the majority of people do not have access to, drawing close similarities to the hierarchy between the scientist and the 'other' (i.e., the object of study). Thus, ADM, like science, allows for oppression of people that are not involved in the development and application of systems.

The second parallel between science and ADM was found in the unconscious use of values. This parallel points to the way in which values enter the process. In science, scientists approach things from their own positionality, which causes them to overlook certain facts or demographics. This has to do with the fact that scientists themselves are shaped by values and their social background: they do not have an objective 'view from nowhere'. If scientists overlook certain factors, these factors will never enter theories. Furthermore, the data that is used to build theories on, often is not representative of society as a whole. As a result, theories and research do not favor each different group equally. We see the same in ADM, where there are many examples of systems that reveal biased judgments against certain groups. Examples that have been used throughout this thesis are the Amazon hiring algorithm and COMPAS. As we have seen, the demographics of AI developers still largely overlap with those of science, causing them to overlook the same factors.

Moreover, the sole endeavor of pursuing science or ADM already reflects a certain value judgment: that this is a worthwhile pursuit. This value plays a significant role in the way we view and use science and ADM, for example in the fact that scientist generally are highly regarded in society. The general valuation of science, ADM and objectivity stays implicit because it is treated as common sense. However, it *is* an example of the value-ladenness of both science and ADM, and we could imagine things to be different. For instance, in countries where stability is more important than innovation, science might be treated with more suspicion because existing ideas could be falsified.

Next, I turned to the differences between science and ADM, which relate to Douglas (2009). It is important to note that the differences between the fields point to an even *stronger* argument against the value-neutrality of ADM, compared to science. In ADM, it is justified to use more (contextual) values, in a more direct way.

The first difference between science and ADM refers to cognitive values. While Douglas' distinction between cognitive and contextual values is useful, cognitive values have a different meaning and purpose in the context of ADM. In science, cognitive values are those values that are used to think through the inferential impact of evidence or data, the risk of making errors, and the impact of research findings (Douglas, 2009). If we would take the same meaning of cognitive values in the context of ADM, it would mean that they are used to think through the impact of a system. However, an important difference between science and ADM is that ADM is inherently much more intertwined with its applications. Whereas scientific theories can be used to explain certain facts in the world, ADM systems generally cannot. ADM's sole purpose is to be applied, to make predictions about new data.

Because application is so important for ADM, one does not only need *cognitive* values to think about the impact a system might make, but also *contextual* values. Generally, it is legitimate to make decisions that are guided by contextual values, such as fairness, equality and inclusion, in order to make a system more applicable. After all, we do not *want* discriminatory systems to be in place. While science needs to stay objective, ADM has to be practical.

The second difference between science and ADM refers to the direct role for values, as described by Douglas (2009). Just as the use of contextual values is more allowed in ADM, it is also more accepted to use a direct role for values. In the direct role for values, values provide stand-alone justifications or reasons to accept a claim. Thus, in this role values take up a role similar to the role evidence normally plays. If values play an indirect role, they do not compete with evidence, but rather determine what should count as enough evidence for a claim. The direct role for values is restricted to very specific circumstances in science. For example, it would be considered bad science if one changes a theory because it might put people in an unfair position.

However, we consciously use values like equality and justice to explicitly guide decisions in ADM. For example, one could choose to use a model design, not because it proved to be very accurate in the past, but because it is more transparent and thus offers more opportunities for people to object against the decisions made by the system. In other words, it is legitimate to favor one system over the other not because it is more *accurate*, but because it is more *fair*.

In Section 4, I have argued that the inherent value-ladenness of ADM does not necessarily mean that ADM is unfair. A system can be unfair because it is biased, but the Amazon hiring example shows that this unfairness is not necessarily caused by a direct role of values. Instead, I have argued that values can be used 'for the better', making ADM fairer and safer. A key condition of this positive use of values is that they are made explicit, so that one can manage them.

The second half of this thesis focused on the third research question: *Where* in the machine learning pipeline do values play a role in ADM, *what* values play a role, and *how* do they play a role? In order to answer this question, I designed a taxonomy of values in ADM. I identified two stages *where* values typically play a role: the development stage and the application stage. Values in the development stage shape the internals of a system, such as the data collection and the choice of model. In the application stage, values are used to justify in what contexts ADM is applied, who employs systems and what role systems play in society. In describing *what* values play a role, I created four broad categories of value judgments in ADM. Computational values (e.g., accuracy) are used to evaluate a system based on its 'objective' consistency with the observed data. Fairness values (e.g., egalitarianism) are used to determine specific fairness criteria in ADM, acting as 'schools of thought' about what is fair. Social values (e.g., equality) are used to support a specific idea about what is (un)fair. Political values (e.g., individualism) are used to justify the application of systems within a particular political climate. Lastly, in discussing *how* values play a role, I distinguished between two roles (inspired by Douglas (2009)): the direct and indirect role for values in ADM. The direct role for values is used when values are directly used to guide decisions about ADM, for example when a model is chosen solely because it is fairer. The indirect role for values is used to weigh the importance, consequences or motivation for a decision. For example, values like accuracy and equality are balanced against each other when deciding which features to include in a model. These values play an indirect role.

The taxonomy can be used as a tool to understand the ways in which a system is or is not neutral. Mapping the value judgments that support a system and its application also helps with recognizing the potential dangers. As objectivity is often associated with neutrality with regard to different (groups of) people, there is a collective belief that objectivity equals unbiasedness (equals fairness). However, as we have seen, both science and ADM are not, and cannot be, objective. It follows that there are biases that need to be tackled, as they can compromise fairness. Thus, by exposing value judgments that play a role in ADM, the taxonomy can be used as a way to deal with bias and the unfair outcomes that can result from this bias. The next section will focus on the relation between the taxonomy and other methods that have been proposed to make ADM fairer.

## 8.2 Relation to other approaches in fair AI

In Section 2, I provided a short overview of the various attempts from the field of AI to make ADM fairer. 'Fair and explainable AI' is a rapidly growing sub-field of AI and is receiving more and more attention from governments and media. This can partly be explained by recent AI missteps, such as the Dutch childcare benefit scandal. I have argued that, while technical approaches to make AI fairer are certainly a step in the right direction, they do not offer an exhaustive solution because they do little to enhance our understanding of the *sources* of bias and unfairness. In general, I distinguished between three main technical approaches in the field of fair AI: understanding bias, mitigating bias, and accounting for bias.

First, approaches to understanding bias mainly focus on comprehending how bias is manifested in data, or investigating how fairness is defined. For example, there are methods to check which features in a data set are most important for a system to base predictions on. This can be used to ensure that these features are actual relevant grounds for a prediction. Furthermore, a significant part of research on fair AI aims at formulating mathematical definitions of fairness, by which we can assess algorithms.

We have seen that, by turning our attention to *values* instead of *bias*, the taxonomy incorporates both these approaches. Acknowledging that data sets are not neutral and understanding which value judgments play a role in them, helps us understand how bias is manifested in data. Analyzing the fairness values that underlie fairness criteria by which a system is evaluated, helps us realize that fairness values are not neutral and can have different outcomes for different people. Thus, focusing on values allows for a comprehensive understanding of how ADM is biased. While most approaches to tackling bias are focused on one or two aspects of bias, the taxonomy covers a multitude of biased factors. This makes the taxonomy a broader approach than many approaches we find in fair AI.

Second, mitigating bias focuses on limiting bias in systems using technical interventions. Mitigation can occur in different stages of the machine learning pipeline. Preprocessing approaches tend to focus on creating representative and balanced data sets. For example, there are techniques to implement computer-made data points in a data set. This can be useful if one group is underrepresented. In-processing approaches focus on changing the internal behavior of a model, in order to create fairer outcomes. For example, models can be made fairer by directly letting mathematical fairness criteria play a role in determining their internal details. This intervention makes AI fairer by directly rejecting the 'thinking steps' of a model that can result in unfair outcomes. Post-processing approaches focus on improving a model after it has been learned from the data. For example, one could alter a model's predictions by promoting or demoting the predictions that are close to the decision boundary. This can result in more equally distributed outcomes.

Interventions to mitigate bias are useful for improving AI, but are limited in two ways when compared to the taxonomy. First, the interventions can be seen as 'temporary fixes', as they do not provide a deeper understanding of *why* bias exists, why it is unfair, and how it relates to values. The taxonomy is designed to be a more sustainable solution by giving a more thorough understanding of these topics and their complex interactions. Second, the interventions focus on improving the internal workings of a system, without looking at the context in which a system is applied (e.g., what is its purpose? Who is subjected to it? What are its possible consequences for individuals and society as a whole?). Instead, the taxonomy also accounts for the 'external world' that is relevant for a system's fairness, by establishing how its application is influenced by value judgments.

Third, accounting for bias focuses on allocating the responsibility for how a system is designed and the possible consequences a system might have. A major trend in this direction is explainable AI, which aims at 'uncovering the black box' of machine learning. These methods focus on the explanation of algorithm outcomes, either in retrospect or by ensuring that models are in itself interpretable enough to follow their 'reasoning'. Making explicit the reasons why a system has come to a decision, makes it easier to recognize if the system is being fair or if it is wrongfully discriminatory. Moreover, having access to explanations of reasoning steps make it easier to object to a decision. However, explainable AI is no guarantee for fair models, as the different aspects of AI (e.g., data collection, internal design, evaluation criteria) do not have to be fair on its own, only explainable or interpretable.

Compared to the taxonomy, explainable AI focuses more on fair procedures for people that are subjected to ADM systems. Instead, the taxonomy reveals the values that are used to shape a system. Thereby, the taxonomy focuses not so much on *how* a model acts, but rather on *why* a model acts the way it does. Furthermore, the taxonomy likely will give less insights into the internal workings of a system (although it does focus on the value judgments that were used in the system design), but instead is more focused on the context in which a system is applied. I believe that this is necessary, because understanding why something is unfair is highly dependent on the context. While approaches in explainable AI might be better at uncovering why a system makes a certain decision, the taxonomy is better at uncovering the value judgments that underlie this decision and indicating why it might be problematic.

Apart from technical interventions, fair AI also focuses on what I call 'contextual solutions': solutions that do not focus on the internals of a system, but rather try to make AI fairer by focusing on understanding the contexts in which systems are applied. Examples of such works are Allhutter et al. (2020) and Crawford & Paglen (2021), who give a thorough analysis of the job seeker algorithm and the manual labeling of images, respectively. These analyses are useful, and also uncover at least one way in which value judgments influence these systems (see Section 2). In this regard, the purpose of their work is similar to the taxonomy. However, these analyses are different from the taxonomy in the sense that they stay limited to *one* (type of) system (i.e., the Austrian unemployment algorithm and image labeling systems). Instead, the taxonomy is purposely designed to be applicable to a wide variety of ADM systems. By offering a framework of guiding questions to be asked about a system, the taxonomy tries to be relevant for a wide range of systems while also offering insights that are specific enough to thoroughly understand the value-ladenness of a system.

Moreover, there are contextual approaches to fair AI that focus on creating fair procedures and regulations. For example, Lepri et al. (2018) emphasize the importance of having diverse teams working on ADM, in order to minimize the risk that teams overlook certain important points or groups. Furthermore, Franzke et al. (2021) propose the DEDA (Data Ethical Decision Aid), which lets developers and deployers think through different questions about a system. The topics of these questions range from privacy and anonymization to transparency and accountability. The DEDA focuses specifically on the 'designing and making' part of ADM. Compared to these methods, the taxonomy is more focused on providing an understanding of a system and its hidden implications on a meta-level, rather than presenting practical tools for fairer ADM. That is why I think that the taxonomy and practical contextual approaches should go hand in hand. Using the taxonomy to think through a system *and*, for example, create a diverse environment in which ADM can be developed, will be more effective than either using either of the two alone.

There is one contextual approach to understanding bias in AI, that comes close to the purpose of the taxonomy of values in this thesis. This approach focuses on understanding the socioeconomic causes of bias that we see in society, as described by Ntoutsi et al. (2020). Institutional bias can be found in many aspects of society and often goes unnoticed. Historical analysis of these biases can help with recognizing prejudice, not only by improving our understanding of which groups are often disadvantaged and why, but also by exposing inequalities that arise from the way data is typically collected. Here, we can see a key similarity with the taxonomy, pointing to the fact that societal values shape AI in a significant way. However, the taxonomy adopts a different approach to make AI fairer. Instead of helping one understand the socioeconomic causes of bias, the taxonomy reveals the connection between *value judgments* and bias. Focusing on the value judgments that underlie a system might touch upon socioeconomic causes of bias as well, but it is not the primary focus. Furthermore, the taxonomy does not only focus on bias in the data (collection), but also in how values play a role in determining the fairness criteria and in justifying the application of a system.

All in all, the mentioned approaches to fair AI are helpful, but tend to focus on a small part of the bigger problem. Approaches either focus on one application, one type of model, or one possible source of bias: the data. I have argued that this is too narrow to fully tackle the problem of unfair AI. Instead, the taxonomy is applicable to a wide range of different applications and models. Moreover, it focuses on different ways in which value judgments shape ADM (thus, not only looking at data). Furthermore, existing approaches mainly focus on improving the internals of a system, while I think focusing on the application is at least as important.

Values and bias are inherently intertwined, because we view bias as personal prejudices that stem from the values one may hold. These values can represent personal beliefs or internalized value judgments that have existed in society throughout history (or a combination of both). However, this connection between values and bias has not been made explicit in past attempts to make AI fairer. By doing so, the taxonomy, and this thesis in general, offers a more thorough understanding of *why* certain things are not neutral or objective, *where* in ADM value judgments can be found, and *how* this can turn out to be unfair for specific individuals or groups. It is important to note that none of the mentioned approaches stand on their own, including the taxonomy I proposed. In order to tackle the problem of biased and unfair ADM systems, approaches should be used to complement each other. Together, they are a step in the right direction towards more powerful, useful and fair AI systems.

# 9 Conclusion

ADM is used to assist decisions that have far-reaching individual or societal implications. Because of its data-driven approach, it is often believed that ADM, like science, is 'objective' or value-neutral. Contrary to humans, ADM systems are thought to have no evolutionary neural biases or value judgments on their own. This should make them more capable of making fair decisions, as values are an important motivator behind bias. The purpose of this thesis is twofold. First, I have demonstrated that ADM systems *do* contain value judgments. Using arguments from philosophy of science, I have shown that ADM, like science, is not and cannot be value-neutral. First, the shared epistemology of both fields is characterized by a 'hunger for objectivity', which in itself reveals a valuation of objectivity and data-driven approaches. Second, in both science and ADM, values often enter the process unconsciously. The positionality and homogeneity of scientists and AI developers often cause them to overlook similar facts or demographics.

A key difference between science and ADM is that value judgments are generally more legitimate in the context of ADM. Whereas science's main goal is to stay objective and neutral, ADM is inherently intertwined with its applications. We actively want values like fairness and equality to play a role in ADM systems. As a result, the boundaries between cognitive and contextual values become blurry in the field of ADM. Furthermore, a direct role for values is much more allowed in ADM: value judgments can play a role similar to evidence in grounding decisions. These differences between science and ADM indicate that ADM is even *more* value-laden than science. This does not mean that all ADM is necessarily unfair, as values can also be used to improve systems.

The second key contribution of this thesis is a taxonomy of values in ADM, in which I have described *where* in the machine learning pipeline values play a role (during development or during application), *what* values play a role (computational, fairness, social, and political values), and *how* they play a role (direct or indirect role). While the taxonomy does provide important insights into the biases and unfairness that results from values, it was not designed to articulate normative judgments about the values in ADM. Thus, the taxonomy does not explicitly answer whether values were used legitimately or illegitimately. However, the insights from applying the taxonomy can of course be used to criticize, modify, or restrict ADM.

The purpose of the taxonomy was illustrated with a case study, focusing on Rotterdam's welfare fraud detection system. Applying the taxonomy revealed that Rotterdam's system made value-laden decisions based on nationality and gender. Furthermore, it became apparent that the system was justified with deeply rooted valuations of crime prevention, economic prosperity, objectivity and efficiency. Applying the taxonomy thus provides us with a more thorough understanding of the potential dangers of a system, by analyzing the grounds on which decisions with regard to the development and application were justified.

Whereas current methods for fair AI tend to focus on technical solutions to tackle bias, the taxonomy provides one with a deeper understanding of the origin and potential consequences of unfairness in ADM. Value judgments, bias, objectivity and unfairness are inherently intertwined. Objectivity is often explained as value-freedom, which explains why one might think that objectivity ensures unbiasedness and fairness. Conversely, value judgments cause bias, which can cause unfairness and stands in the way of objectivity.

This taxonomy is not the answer to unfair ADM, and it should be used to complement existing methods for fair AI. As AI and ADM have become omnipresent tools in many layers of society, no efforts should be spared to make their workings, outcomes, procedures and applications fairer for all.

## Acknowledgments

First and foremost, I would like to express my gratitude towards my supervisor, Dr. Dominik Klein, who has provided me with great feedback throughout the project. Our discussions were as much informative as they were fun, and have definitely contributed to the quality (and quantity) of this thesis. Furthermore, I am thankful to my second examiner, Dr. Michael De, for taking the time to read and assess this thesis. Finally, I would like to thank my family and friends for their unconditional support and feedback during the past months, especially my parents, Rebecca, Aafje and Ellen. Being surrounded by such strong and intelligent women is something that inspires me deeply.

## References

Adam, A. (1995). Artificial Intelligence and Women's Knowledge: What can feminist epistemologies tell us? *Women's Studies International Forum*, *18*(4), 407–415.

Allhutter, D., Cech, F., Fischer, F., Grill, G., & Mager, A. (2020). Algorithmic Profiling of Job Seekers in Austria: How austerity politics are made effective. *Frontiers in Big Data*, *3*, 5.

Alon-Barkat, S., & Busuioc, M. (2023). Human–AI Interactions in Public Sector Decision Making: "Automation bias" and "selective adherence" to algorithmic advice. *Journal of Public Administration Research and Theory*, *33*(1), 153–169.

Anderson, E. (2004). Uses of Value Judgments in Science: A general argument, with lessons from a case study of feminist research on divorce. *Hypatia*, *19*(1), 1–24.

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016). *Machine Bias.* Retrieved December 2022, from https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing/

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

van Bekkum, M., & Borgesius, F. Z. (2021). Digital Welfare Fraud Detection and the Dutch SyRI Judgment. *European Journal of Social Security*, *23*(4), 323–340.

Binns, R. (2018). What Can Political Philosophy Teach Us About Algorithmic Fairness? *IEEE Security & Privacy*, *16*(3), 73–80.

Bobel, C., Winkler, I. T., Fahs, B., Hasson, K. A., Kissling, E. A., & Roberts, T.-A. (2020). *The Palgrave Handbook of Critical Menstruation Studies.* Singapore: Springer Nature.

Braun, J.-C., Constantaras, E., Aung, H., Geiger, G., Mehrotra, D., & Howden, D. (2023). *Suspicion Machines Methodology.* Retrieved April 2023, from https://www.lighthousereports.com/suspicion-machines-methodology/

Buolamwini, J., & Gebru, T. (2018). Gender Shades: Intersectional accuracy disparities in commercial gender classification. *Conference on Fairness, Accountability and Transparency*, 77–91.

Burgess, M., Schot, E., & Geiger, G. (2023). *This Algorithm Could Ruin Your Life.* Retrieved April 2023, from https://www.wired.co.uk/article/welfare-algorithms-discrimination

College voor de Rechten van de Mens. (2021). *Discriminatie door Risicoprofielen: Een mensenrechtelijk toetsingskader.* 's-Gravenzande: Van Deventer.

Constantaras, E., Geiger, G., Braun, J.-C., & Aung, H. (2023). *Inside the Suspicion Machine.* Retrieved April 2023, from `https://www.wired.com/story/welfare-state-algorithms/`

Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., & Huq, A. (2017). Algorithmic Decision Making and the Cost of Fairness. *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 797–806.

Crasnow, S. (2004). Objectivity: Feminism, values, and science. *Hypatia*, *19*(1), 280–291.

Crawford, K., & Paglen, T. (2021). Excavating AI: The politics of images in machine learning training sets. *AI & Society*, 1–12.

D'Amour, A., Srinivasan, H., Atwood, J., Baljekar, P., Sculley, D., & Halpern, Y. (2020). Fairness is Not Static: Deeper understanding of long term fairness via simulation studies. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 525–534.

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

Dastin, J. (2018, Oct). *Amazon Scraps Secret AI Recruiting Tool That Showed Bias Against Women.* Thomson Reuters. Retrieved January 2023, from `https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G`

Douglas, H. (2009). *Science, Policy, and the Value-Free Ideal.* Pittsburgh, Pennsylvania: University of Pittsburgh Press.

Fiske, S. T. (1998). Stereotyping, Prejudice, and Discrimination. In D. Gilbert, S. Fiske, & G. Lindzey (Eds.), *Handbook of Social Psychology* (pp. 357–411). Boston: McGraw-Hill.

Franzke, A. S., Muis, I., & Schäfer, M. T. (2021). Data Ethics Decision Aid (DEDA): A dialogical framework for ethical inquiry of ai and data projects in The Netherlands. *Ethics and Information Technology*, 1–17.

Friedler, S. A., Scheidegger, C., & Venkatasubramanian, S. (2016). On the (Im)Possibility of Fairness. *arXiv preprint arXiv:1609.07236*.

Halpin, Z. T. (1989). Scientific Objectivity and the Concept of "The Other". *Women's Studies International Forum*, *12*(3), 285–294.

Hedden, B. (2021). On Statistical Criteria of Algorithmic Fairness. *Philosophy and Public Affairs*, *49*(2), 209–231.

Kamiran, F., & Calders, T. (2009). Classifying Without Discriminating. *2009 2nd International Conference on Computer, Control and Communication*, 1–6.

Kamiran, F., Mansha, S., Karim, A., & Zhang, X. (2018). Exploiting Reject Option in Classification for Social Discrimination Control. *Information Sciences*, *425*, 18–33.

Kuhn, T. S. (1977). Objectivity, Value Judgment, and Theory Choice. *Arguing about Science*, 74–86.

Kusner, M. J., Loftus, J., Russell, C., & Silva, R. (2017). Counterfactual Fairness. *Advances in Neural Information Processing Systems*, *30*.

Lacey, H. (1999). *Is Science Value Free? Values and Scientific Understanding.* London: Routledge.

Lacey, H. (2018). Roles for Values in Scientific Activities. *Axiomathes*, *28*(6), 603–618.

Langenkamp, M., Costa, A., & Cheung, C. (2020). Hiring Fairly in the Age of Algorithms. *arXiv preprint arXiv:2004.07132*.

Lepri, B., Oliver, N., Letouzé, E., Pentland, A., & Vinck, P. (2018). Fair, Transparent, and Accountable Algorithmic Decision-Making Processes: The premise, the proposed solutions, and the open challenges. *Philosophy & Technology*, *31*, 611–627.

Longino, H. E. (1995). Gender, Politics, and the Theoretical Virtues. *Synthese*, *104*(3), 383–397.

Longino, H. E. (1996). Cognitive and Non-Cognitive Values in Science: Rethinking the dichotomy. In L. H. Nelson & J. Nelson (Eds.), *Feminism, Science, and the Philosophy of Science* (pp. 39–58). Dordrecht: Springer Netherlands.

McQuillan, D. (2018). Data Science as Machinic Neoplatonism. *Philosophy & Technology*, *31*(2), 253–272.

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*, *8*(1), 141–163.

Molenberghs, P. (2013). The Neuroscience of In-Group Bias. *Neuroscience & Biobehavioral Reviews*, *37*(8), 1530–1536.

Moreau, S. (2020). *Faces of Inequality: A theory of wrongful discrimination.* Oxford: Oxford University Press.

Ntoutsi, E., Fafalios, P., Gadiraju, U., Iosifidis, V., Nejdl, W., Vidal, M.-E., … Staab, S. (2020). Bias in Data-Driven Artificial Intelligence Systems—An introductory survey. *WIREs Data Mining and Knowledge Discovery*, *10*(3).

Perez, C. C. (2019). *Invisible Women: Data bias in a world designed for men.* New York, NY: Abrams Press.

Poursabzi-Sangdeh, F., Goldstein, D. G., Hofman, J. M., Wortman Vaughan, J. W., & Wallach, H. (2021). Manipulating and Measuring Model Interpretability. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–52.

Reiss, J., & Sprenger, J. (2020). Scientific Objectivity. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Winter 2020 ed.). Metaphysics Research Lab, Stanford University. `https://plato.stanford.edu/archives/win2020/entries/scientific-objectivity/`.

Rieder, G., & Simon, J. (2016). Datatrust: Or, the political quest for numerical evidence and the epistemologies of Big Data. *Big Data & Society*, *3*(1), 1–6.

Russell, S. J., & Norvig, P. (2010). *Artificial Intelligence: A modern approach* (3rd ed.). Saddle River, NJ: Prentice Hall.

van Schendel, S. (2019). The Challenges of Risk Profiling Used by Law Enforcement: Examining the cases of COMPAS and SyRI. In L. Reins (Ed.), *Regulating New Technologies in Uncertain Times* (pp. 225–240). The Hague: T.M.C. Asser Press.

Selbst, A. D., Boyd, D., Friedler, S. A., Venkatasubramanian, S., & Vertesi, J. (2019). Fairness and Abstraction in Sociotechnical Systems. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 59–68.

Selvaraju, R. R., Das, A., Vedantam, R., Cogswell, M., Parikh, D., & Batra, D. (2016). Grad-CAM: Why did you say that? *arXiv preprint arXiv:1611.07450*.

Verma, S., & Rubin, J. (2018). Fairness Definitions Explained. *2018 IEEE/ACM International Workshop on Software Fairness (FairWare)*, 1–7.

Wellner, G., & Rothman, T. (2020). Feminist AI: Can we expect our AI systems to become feminist? *Philosophy & Technology*, *33*(2), 191–205.

West, J., & Bhattacharya, M. (2016). Intelligent Financial Fraud Detection: A comprehensive review. *Computers & security*, *57*, 47–66.

Winterson, J. (2021). *12 Bytes: How we got here, where we might go next.* London: Jonathan Cape.

Zou, J., & Schiebinger, L. (2018). AI Can Be Sexist and Racist — It's time to make it fair. *Nature*, *559*, 324–326.